

Handbook of  
**ECONOMIC FIELD  
EXPERIMENTS**

Handbook of Field Experiments

VOLUME **2**

# Handbook of **ECONOMIC FIELD EXPERIMENTS**

Handbook of Field Experiments

VOLUME **2**

Edited by

**ABHIJIT VINAYAK BANERJEE**

Massachusetts Institute of Technology  
Cambridge, MA, United States

**ESTHER DUFLO**

Massachusetts Institute of Technology  
Cambridge, MA, United States



**North-Holland**

An imprint of Elsevier

North-Holland is an imprint of Elsevier  
Radarweg 29, PO Box 211, 1000 AE Amsterdam, Netherlands  
The Boulevard, Langford Lane, Kidlington, Oxford OX5 1GB, United Kingdom

Copyright © 2017 Elsevier B.V. All rights reserved.

No part of this publication may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, recording, or any information storage and retrieval system, without permission in writing from the publisher. Details on how to seek permission, further information about the Publisher's permissions policies and our arrangements with organizations such as the Copyright Clearance Center and the Copyright Licensing Agency, can be found at our website: [www.elsevier.com/permissions](http://www.elsevier.com/permissions).

This book and the individual contributions contained in it are protected under copyright by the Publisher (other than as may be noted herein).

## Notices

Knowledge and best practice in this field are constantly changing. As new research and experience broaden our understanding, changes in research methods, professional practices, or medical treatment may become necessary.

Practitioners and researchers must always rely on their own experience and knowledge in evaluating and using any information, methods, compounds, or experiments described herein. In using such information or methods they should be mindful of their own safety and the safety of others, including parties for whom they have a professional responsibility.

To the fullest extent of the law, neither the Publisher nor the authors, contributors, or editors, assume any liability for any injury and/or damage to persons or property as a matter of products liability, negligence or otherwise, or from any use or operation of any methods, products, instructions, or ideas contained in the material herein.

ISBN: 978-0-444-64011-6

ISSN: 2214-658X

For information on all Elsevier publications visit our website at <https://www.elsevier.com/books-and-journals>



Working together  
to grow libraries in  
developing countries

[www.elsevier.com](http://www.elsevier.com) • [www.bookaid.org](http://www.bookaid.org)

*Publisher:* Zoe Kruze

*Acquisition Editor:* Kirsten Shankland

*Editorial Project Manager:* Edward Payne

*Production Project Manager:* Stalin Viswanathan

*Designer:* Mark Rogers

Typeset by TNQ Books and Journals

## INTRODUCTION TO THE SERIES

The aim of the *Handbooks in Economics* series is to produce Handbooks for various branches of economics, each of which is a definitive source, reference, and teaching supplement for use by professional researchers and advanced graduate students. Each Handbook provides self-contained surveys of the current state of a branch of economics in the form of chapters prepared by leading specialists on various aspects of this branch of economics. These surveys summarize not only received results but also newer developments, from recent journal articles and discussion papers. Some original material is also included, but the main goal is to provide comprehensive and accessible surveys. The Handbooks are intended to provide not only useful reference volumes for professional collections but also possible supplementary readings for advanced courses for graduate students in economics.

**Founding Editors**  
**Kenneth J. Arrow and Michael D. Intriligator**

# CONTRIBUTORS

## Volume 1

### **O. Al-Ubaydli**

Bahrain Center for Strategic, International and Energy Studies, Manama, Bahrain; George Mason University, Fairfax, VA, United States; Mercatus Center, Arlington, VA, United States

### **S. Athey**

Stanford University, Stanford, CA, United States; NBER (National Bureau of Economic Research), Cambridge, MA, United States

### **A.V. Banerjee**

Massachusetts Institute of Technology, Cambridge, MA, United States; NBER (National Bureau of Economic Research), Cambridge, MA, United States

### **M. Bertrand**

University of Chicago Booth School of Business, Chicago, IL, United States

### **S. Chassang**

New York University, New York, NY, United States

### **E. Duflo**

Massachusetts Institute of Technology, Cambridge, MA, United States

### **A.S. Gerber**

Yale University, New Haven, CT, United States

### **R. Glennerster**

Massachusetts Institute of Technology, J-PAL, Cambridge, MA, United States

### **U. Gneezy**

University of California, San Diego, La Jolla, CA, United States; University of Amsterdam, Amsterdam, Netherlands

### **D.P. Green**

Columbia University, New York, NY, United States

### **J.M. Gueron**

President Emerita, MDRC, New York, NY, United States

### **A. Imas**

Carnegie Mellon University, Pittsburgh, PA, United States

### **G.W. Imbens**

Stanford University, Stanford, CA, United States; NBER (National Bureau of Economic Research), Cambridge, MA, United States

**J.A. List**

University of Chicago, Chicago, IL, United States; NBER (National Bureau of Economic Research), Cambridge, MA, United States

**E.L. Paluck**

Princeton University, Princeton, NJ, United States

**D. Simester**

MIT Sloan School of Management, Cambridge, MA, United States

**E. Shafir**

Princeton University, Princeton, NJ, United States

**E. Snowberg**

California Institute of Technology, Pasadena, CA, United States; NBER (National Bureau of Economic Research), Cambridge, MA, United States

## Volume 2

**W.J. Congdon**

ideas42, New York, NY, United States

**A. de Janvry**

University of California, Berkeley, Berkeley, CA, United States

**P. Dupas**

Stanford University, Stanford, CA, United States; NBER (National Bureau of Economic Research), Cambridge, MA, United States; Center for Education Policy Research, Cambridge, MA, United States

**F. Finan**

University of California, Berkeley, Berkeley, CA, United States

**R.G. Fryer, Jr.**

Harvard University, Cambridge, MA, United States; NBER (National Bureau of Economic Research), Cambridge, MA, United States

**R. Hanna**

Harvard University, Cambridge, MA, United States

**D. Karlan**

Yale University, New Haven, CT, United States

**J.R. Kling**

Congressional Budget Office, Washington, DC, United States; NBER (National Bureau of Economic Research), Cambridge, MA, United States

**J. Ludwig**

NBER (National Bureau of Economic Research), Cambridge, MA, United States; University of Chicago, Chicago, IL, United States

**E. Miguel**

University of California, Berkeley, Berkeley, CA, United States; NBER (National Bureau of Economic Research), Cambridge, MA, United States

**S. Mullainathan**

NBER (National Bureau of Economic Research), Cambridge, MA, United States; Harvard University, Cambridge, MA, United States

**K. Muralidharan**

University of California, San Diego, La Jolla, CA, United States; NBER (National Bureau of Economic Research), Cambridge, MA, United States; Jameel Poverty Action Lab, Cambridge, MA, United States

**B.A. Olken**

Massachusetts Institute of Technology, Cambridge, MA, United States

**R. Pande**

Harvard University, Cambridge, MA, United States

**J. Rothstein**

University of California, Berkeley, Berkeley, CA, United States; NBER (National Bureau of Economic Research), Cambridge, MA, United States

**E. Sadoulet**

University of California, Berkeley, Berkeley, CA, United States

**T. Suri**

MIT Sloan School of Management, Cambridge, MA, United States

**T. von Wachter**

NBER (National Bureau of Economic Research), Cambridge, MA, United States; University of California, Los Angeles, Los Angeles, CA, United States

# CHAPTER 1

## Impacts and Determinants of Health Levels in Low-Income Countries

P. Dupas\*, §¶, 1, E. Miguel ||§

\*Stanford University, Stanford, CA, United States

§NBER (National Bureau of Economic Research), Cambridge, MA, United States

¶Center for Education Policy Research, Cambridge, MA, United States

||University of California, Berkeley, Berkeley, CA, United States

<sup>1</sup>Corresponding author: E-mail: pdupas@stanford.edu

### Contents

1. Introduction	4
2. Methodological Section	6
2.1 Experimenting to estimate impacts of health improvements: beware of externalities	8
2.2 Experimenting to understand the determinants of health behavior: beware of measurement	14
2.3 Research transparency, registration, and preanalysis plans	16
3. Experimental Estimates of the Impact of Health on Individual Productivity	18
3.1 Impacts of adult health and nutrition on productivity	18
3.2 Impacts of child health and nutrition on education	23
3.3 Impacts of child health and nutrition on later outcomes	28
4. Environmental/Infrastructural Determinants of Health	31
5. Demand for Health Products and Healthcare	34
5.1 Pricing experiments	36
5.1.1 Methods to estimate the demand curve	36
5.1.2 Results of pricing experiments	40
5.2 Liquidity experiments: credit and cash transfer experiments	47
5.3 Information experiments	49
5.3.1 Impact of information on willingness to pay	49
5.3.2 Impact of information on health behavior change	51
5.3.3 Impact of tailored information on behavior change	54
5.3.4 Impact of targeted information	57
5.4 Schooling experiments	57
5.5 Nonmonetary cost experiments	58
5.6 Incentive experiments	61
5.7 Psychology experiments	62
5.8 Taking stock: how important is the role of present bias in explaining observed preventive health behaviors?	64
5.9 Commitment experiments	67
6. Supply of Health Care	69
6.1 Experimental audit studies	69
6.2 Monitoring experiments	72
6.2.1 District-level contracting	73

6.2.2 <i>Top-down, input-based incentives</i>	74
6.2.3 <i>Top-down, output-based incentives</i>	75
6.2.4 <i>Bottom-up: beneficiary oversight</i>	77
<b>6.3 Improving the quality of informal providers</b>	<b>79</b>
<b>7. Conclusion</b>	<b>82</b>
<b>References</b>	<b>84</b>

## Abstract

Improved health in low-income countries could considerably improve wellbeing and possibly promote economic growth. The last decade has seen a surge in field experiments designed to understand the barriers that households and governments face in investing in health and how these barriers can be overcome, and to assess the impacts of subsequent health gains. This chapter first discusses the methodological pitfalls that field experiments in the health sector are particularly susceptible to, then reviews the evidence that rigorous field experiments have generated so far. While the link from in utero and child health to later outcomes has increasingly been established, few experiments have estimated the impacts of health on contemporaneous productivity among adults, and few experiments have explored the potential for infrastructural programs to impact health outcomes. Many more studies have examined the determinants of individual health behavior, on the side of consumers as well as among providers of health products and services.

## Keywords

Behavior; Epidemiology; Externalities; Incentives; Information; Prevention; Public health; Subsidies

## JEL Codes

C93: Field Experiments; D12: Consumer Economics: Empirical Analysis; I1: Health; I12: Health Behavior; I15: Health and Economic Development; O12: Microeconomic Analyses of Economic Development

## 1. INTRODUCTION

The links between health and economic development are many and varied. One of the most robust stylized facts of economic development is that higher income levels correlate strongly with longer life spans, lower infant mortality, and reduced illness throughout the life course ([Deaton, 2013](#)). Infectious diseases that kill millions in poor countries are largely unknown in the world's wealthy societies, while sophisticated new curative technologies, procedures, and pharmaceutical advances often appear in wealthy economies years before they are available in low-income regions. The recent ravages of HIV/AIDS in Sub-Saharan Africa—which has killed tens of millions, and counting—have only deepened the divide between the global health haves and have-nots. Even the briefest introspection makes it obvious that health levels are critical determinants of human wellbeing even beyond their impact on economic productivity, and that the health gaps across countries are a major contributor to global inequities.

This is a powerful and well-known pattern, but its underlying causes are not obvious or entirely clear. There are many channels that could plausibly contribute to the link between

wealth and health. On the one hand, higher incomes may allow individuals, households and whole societies to invest more resources in better health prevention and treatment (as well as better nutrition), leading to improved health outcomes. On the other, individuals with poor health may not be able to work as long, as hard, or as effectively as their healthier peers, leading to lower incomes and living standards for themselves, and potentially for their households and offspring. Adding further complexity is the possibility that both economic outcomes and health levels might be jointly determined by a third factor, such as the effectiveness of government public good provision—which itself might be a function of the quality of government institutions—or the design of foreign aid programs. All of these mechanisms also potentially interact in complex ways.

The research community has worked over the past few decades to better understand each of these different potential relationships and mechanisms, and this survey is an attempt to draw together the evidence from experimental studies on these important topics. Making sense of the bidirectional relationship between health and income is critical for both our scholarly understanding of the world in which we live, and for the effective design of public policies. As we discuss in this chapter, the evidence base is growing on the impacts and determinants of health levels in developing countries, but there remain many knowledge gaps, and we highlight areas where further research would be useful. In doing so, we build on the many recent surveys and chapters that have tackled related issues.

One fundamental relationship that has been the focus of many studies is the causal link between improved health and economic productivity. This is actually a challenging relationship to estimate rigorously, given the bidirectional relationships described above. Several recent experimental studies have begun to make progress, and have shown large impacts of gains in both in utero and child health on later educational and economic outcomes ([Glewwe and Miguel, 2008](#)). With a handful of notable exceptions, there is limited experimental evidence tracing links from child health gains to adult economic outcomes due to the paucity of long-term panel datasets in low-income settings (although this is starting to change with more sophisticated longitudinal data collection approaches). In the meantime, some of the best evidence in this area will be have to come from natural experiments or other nonexperimental designs ([Bleakley, 2010a](#)). There is even less evidence on links between health levels and contemporaneous adult productivity. This is harder to estimate than it might appear in part due to difficulties both in intervening to improve adult health, as well as in measuring individual output in many economic occupations (e.g., household subsistence farming, self-employment and informal work) in poor countries, although a new body of evidence is attempting to overcome these challenges.

There has been much more high-quality experimental research on the demand for health, with many studies documenting that even poor households in developing countries spend large sums on acute health care. However, a growing body of evidence

indicates that demand for many preventive technologies, as well as a range of new and seemingly useful health products, is far lower than might seem optimal ([Dupas, 2011a](#)).

There has recently been active debate over the right “model” of health behavior among the poor, with an excellent recent survey ([Kremer and Glennerster, 2011](#)) arguing that the low demand for prevention is consistent with the importance of present biased preferences for many individuals. We assess the patterns in the exploding literature in this area, describe evidence that we believe is consistent with this interpretation as well as other patterns that we believe may be more consistent with alternative theoretical models and interpretations. Overall, our view on the role of present bias is more nuanced and uncertain than [Kremer and Glennerster \(2011\)](#).

There is a growing consensus that supply-side constraints, both in terms of access and the quality of health provision, could be a major determinant of poor health outcomes in developing countries ([Das and Hammer, 2014](#)). The lack of high quality health care in many settings is important in its own right, and also affects interpretation of other findings in this literature. The lack of high quality service options could be a key driver of the low demand for healthcare observed in many settings (other than in emergency curative situations, where the consequences of nontreatment are especially severe and immediate). The generally low quality of health provision in poor countries also makes it more challenging to study the link between health and economic productivity, to the extent that even large-scale health policy reforms do not “deliver” better population health. Experimental studies that aim to improve individual health by working through existing health institutions may be stymied by the limited capacity of those institutions.

Throughout this chapter, we survey the high-quality evidence from developing countries on these topics, highlight holes in the existing literature, and point the way forward for future research in the area. Before we survey the evidence, we review some key methodological issues in [Section 2](#) that pertain to field experiments in health and that are critical to reading the evidence. [Section 3](#) surveys experimental estimates of the impact of health on individual productivity, including studies in child as well as adult interventions. [Section 4](#) surveys recent experimental studies on the environmental determinants of health, including clean water interventions. [Section 5](#) discusses the exploding literature on the demand for health and healthcare, while [Section 6](#) surveys work on the supply of health care. The final section concludes.

## **2. METHODOLOGICAL SECTION**

Experiments in the health sector have been prominent among the field experiments carried out in development economics over the past two decades, and they have highlighted a number of important methodological issues related to the estimation of externalities, variable measurement, and preregistration and research transparency. We discuss each of these issues in turn in the three subsections that follow.

Before launching into the detailed discussion, a few observations about the differences between field experiments conducted by development economists and those carried out among clinical trialists and epidemiologists are in order. One key difference relates to the goals of their studies, in relation to the widely used distinction between *efficacy trials* and *efficiency trials*. Efficacy trials are designed to capture the impact of an intervention under the most controlled and ideal circumstances possible, while efficiency trials capture effects under more authentic real-world conditions ([Singhal et al., 2014](#)).

In reality, many studies lie somewhere in between these two extremes, with both partial study control and some degree of realism. While medical researchers typically carry out both types of studies, and often make a sharp distinction between the two, most recent field experiments in economics have tended to be closer to efficiency trials. Many of these studies have been carried out in close collaboration with government or nongovernmental organization (NGO) partners, and have been implemented as “real” programs, rather than experiments that are carried out directly by the researchers themselves, as in many efficacy trials.

Even more importantly, there is a major difference in the role of theory in the design of experiments by economists versus health researchers. Economists (and other social scientists) often design experiments to shed light on underlying theoretical mechanisms, to inform ongoing theoretical debates, and measure and estimate endogenous behavioral responses. These behavioral responses may shed light on broader issues beyond the experimental intervention at hand, and thus could contribute to greater external validity of the results.

This distinction between the types of studies carried out by medical researchers versus development economists working on health topics has a number of important implications that will become apparent in the course of this chapter. One has to do with the quality “standards” and perceptions of the “risk of bias” in a particular design. For medical trialists accustomed to the CONSORT standards or other medical efficacy trial reporting guidelines, studies that do not feature double-blinding, and thus run the “risk” of endogenous behavioral responses to the medical intervention, are considered less reliable than those studies that employ double-blinding (for a detailed discussion, see [Eble et al., 2015](#)). While a few of the studies conducted by economists surveyed below do feature double-blinding (most notably [Thomas et al., 2003, 2006](#)), in nearly all settings blinding participants to their status is either logically difficult (for instance, if government partners are unwilling to distribute placebo treatments to some of their population) or even impossible.

To illustrate, how would you provide a placebo treatment in a study investigating the impact of the distribution of antimalarial bed nets? Even in settings that might seem promising for placebo treatments, such as the community-level deworming treatments discussed below at several points, blinding participants to their status is basically impossible. Deworming generates side effects (mainly gastrointestinal distress) in roughly

10% of those who take the pills, so community members in a placebo community would quickly deduce that they were in fact not receiving real deworming drugs if there are few or no cases of side effects.

As noted above, endogenous behavioral responses are often exactly what we economists (and other social scientists) set out to measure and estimate in our field experiments, and thus are to be embraced rather than rejected as symptomatic of a “low quality” research design that is at “high risk of bias.” Finally, economists’ interest in many cases in the cost-effectiveness, economic returns, or fiscal implications of particular real-world health interventions once again make efficacy trials of inherently less interest in most cases than more realistic effectiveness trial approaches. In fact, the differences in outcomes between efficacy trials and effectiveness trials are of great interest to social sciences, since understanding why an intervention that “works” in a highly controlled settings might “fail” in a more realistic setting can shed light on the functioning—and limitations—of existing organizations and institutions.

Taken together, it is clear to us that the experimental literature on health interventions in economics (and increasingly in other social sciences such as political science) often has very different objectives than medical, public health and epidemiological research, and thus different methodologies are often called for. Researchers working on health topics in development economics have not simply been able to import existing medical trial methods wholesale, but have instead been quite innovative in developing new approaches to estimating externalities, in measurement, and regarding issues of preregistration and transparency, as discussed in the three subsections below.

## **2.1 Experimenting to estimate impacts of health improvements: beware of externalities**

Field experiments in development economics focusing on the health sector have been innovative in adapting and creating new approaches to estimating treatment externalities. Treatment externalities go by many different names in different subfields and disciplines—including spillovers, contamination, herd immunity, and indirect effects, among others—and they have been of interest to statisticians and epidemiologists for a long time (for classic treatments, refer to [Cox, 1958](#); [Rubin, 1990](#); [Fine, 1993](#); [Rosenbaum, 2007](#)). However, despite their theoretical importance for the health field, especially in low-income societies where infectious diseases account for a large share of the disease burden, these issues have received far less empirical attention in public health and epidemiological research.

As surveyed in a recent synthetic review ([Benjamin-Chung et al., 2015](#)), the rapid growth in empirical studies of treatment externalities in epidemiology began after 2000, at roughly the same time that such empirical treatments became more common in economics, although there were a handful of earlier empirical treatments of the issue (for instance, [Paul et al., 1962](#); [Cooper and Fitch, 1983](#)). This literature has tended to

focus on low income settings in Asia, Latin America and Africa in both the public health literature and in economics. As [Benjamin-Chung et al. \(2015\)](#) show in their detailed review of the existing literature, both these early public health studies, as well as most recent treatments (for instance, [Forleo-Neto et al., 1999](#); [Ali et al., 2005, 2008, 2013](#)), tend to focus on “herd immunity” effects in vaccine treatment programs, and they estimate treatment spillovers using the “treatment coverage mean,” i.e., the proportion of individuals in the area who received treatment, as the key measure of exposure. They then examine whether there is a lower risk of later infection in sites where more people received treatment relative to areas where fewer people were treated. These studies provide consistent support for the existence of positive vaccine spillover benefits among those who did not receive the vaccine themselves.

This is an empirically sensible approach but it has a number of immediate limitations. First, many studies using these approaches typically leave out the question of how and why particular sites had lower vaccine coverage than other areas unanswered. This is potentially problematic to the extent that coverage rates are affected by omitted variables (“confounders”) such as the local disease environment, capacity of local health institutions, and perhaps local attitudes towards particular diseases, all of which could both affect coverage as well as population health outcomes. For instance, it is plausible that areas where populations have less awareness of, or support for, treatment of a disease might both have lower coverage rates and higher subsequent infection (although there are other possible confounders that would lead to bias in the opposite direction). Second, this approach is typically quite imprecise about the geographic area that is relevant for spillovers; the choice of geographic area often appears quite ad hoc; and different studies use different definitions of a community or site, thus leading to a lack of comparability across studies. Taken together, this implies that the evidence base within public health regarding the extent of treatment spillovers is not extremely solid, and moreover, the evidence generated so far has tended to focus on a narrow set of treatments, namely, vaccinations.

While it may be surprising that there has not been more empirical research on the empirical estimation of spillover effects within public health and epidemiology (despite the theoretical centrality of these issues in these fields, i.e., they are even in the name “epidemiology”), we speculate that it may be the result of a tendency in most empirical health studies to focus on “standard” RCT empirical approaches that compare treatment to control groups directly, and that tend to regard any sort of externality effect as a source of “contamination” that is to be avoided or minimized, rather than as a key element of our understanding of the overall treatment effect. Economists who have worked on these topics in health have instead been more open to embracing the estimation of externalities, perhaps in part because norms regarding the “right” way to carry out field experiments are less established in economics (given how recently these tools have been adopted in the field), and also given the importance of spillover effects within public

economics theory to potentially rationalize public subsidies for health treatments and interventions ([Dybvig and Spatt, 1983](#)).

The [Miguel and Kremer \(2004\)](#) paper on school-based deworming impacts in Kenya is among the first health studies in development economics to experimentally estimate externality effects. In their main analysis, they exploit the variation in deworming treatment status generated by the experimental assignment of schools to early versus late treatment (in a phase-in, or stepped wedge, research design), and show that this generates extensive variation in local “exposure” to treated schools within 3 km and up to 6 km away from sample schools. Their econometric approach conditions on the total density of local school pupil population within a particular geographic distance, and conditioning on this quantity, the experimental design implies that the number of treated schools is experimentally assigned and should thus be orthogonal to other local observables and unobservables. They cannot reject that the observed characteristics of schools with lots of “exposure” to other local treatment schools are the same as for schools with little such exposure. Thus this analytical approach—which [Benjamin-Chung et al. \(2015\)](#) terms “estimation of spillovers conditional on treatment density”—addresses both of the limitations of most existing empirical research on spillovers in public health described above. In particular, the assignment of exposure to treatment spillovers is assigned experimentally (and thus should be largely free of the possible omitted variable bias, or confounding, that could affect most existing vaccine studies in epidemiology), and this approach also makes precise the extent of externalities within precisely defined geographic distances away from a particular site.

Formally, [Miguel and Kremer \(2004\)](#) estimate the following econometric model:

$$Y_{ijt} = \alpha + \beta T_{it} + X'_{ijt} \delta + \sum_d (\gamma_d \cdot N_{dit}^T) + \sum_d (\phi d \cdot N_{dit}) + e_{ijt}. \quad (1)$$

$Y_{ijt}$  is the individual health or education outcome, where  $i$  refers to the school,  $j$  to the student, and  $t$  the time period;  $T_{it}$  is an indicator variable for school assignment to deworming treatment; and  $X_{ijt}$  are school and pupil characteristics, and time controls.  $N_{dit}$  is the total number of pupils in primary schools at distance  $d$  from school  $i$  in year  $t$ , and  $N_{dit}^T$  is the number of these pupils in schools randomly assigned to deworming treatment. In their example,  $d = 03$  denotes schools that are located within 3 km of school  $i$ , and  $d = 36$  denotes schools that are located between 3 and 6 km away. Individual disturbance terms were assumed to be independent across schools, but are allowed to be correlated for observations within the same school. In this framework  $\beta + \sum_d (\gamma_d \bar{N}_{dit}^T)$  is the average effect of deworming treatment on overall infection prevalence in treatment schools, where  $\bar{N}_{dit}^T$  is the average number of treatment school pupils located at distance  $d$  from the school. Under spatial externality models in which a reduction in worm prevalence at one school affects neighboring schools and this in turn affects their neighbors,

some externalities would spill over to even greater distances, in which case Eq. (1) yields a lower bound on treatment effects, a point that Baird et al. (2015) show formally.  $\beta$  captures both direct effects of deworming treatment on the treated, as well as any externalities on untreated pupils within the treatment schools.

Miguel and Kremer (2004) also use another source of variation to estimate spillover effects within treated school communities. Within communities, a subset of the population was not assigned to treatment, namely, older girls for whom the deworming drugs were considered potentially dangerous at the time of the original study (due to potential embryotoxicity), and other students simply did not receive treatment, usually because they did not attend school on the announced day of treatment or did not receive parental consent for treatment. The comparison of subsequent infection rates among those in treatment schools who did not themselves take deworming drugs, compared to those who did take the drugs, is potentially problematic due to nonrandom selection into treatment, and any such differences lack a reliable counterfactual (since time trends or other secular changes might affect both groups).

However, Miguel and Kremer (2004) are able to exploit the fact that the same treatment inclusion rules were used in subsequent years of the program as later treatment groups were phased into deworming, and they compare infection rates among those who did not receive deworming treatment when it was available in their school, to those in other schools who were not yet offered deworming but who we know did not receive treatment when later offered the opportunity. This approach potentially addresses much of the “selection” into deworming treatment, as long as patterns of selection remain roughly constant across years 1 and 2 of the study. This is a “within-cluster spillover effect,” and when focusing on the excluded older girls, Benjamin-Chung et al. (2015) term it a “within-cluster spillover effect among ineligibles.”

There is evidence for large and statistically significant externality effects on both worm infection rates, and on subsequent school participation rates, using both sources of variation in Miguel and Kremer (2004), namely, the spillover estimates conditional on local treatment density, and the within-cluster spillover effect. These effects are concentrated within school communities, and extend out to at least 3 km away from treatment schools.

In a follow-up study in the same area of Kenya, Ozier (2014) also generates within-cluster spillover effect estimates among ineligible, by focusing on children who were 0–2 years old when the program was launched, and thus were too young to have directly received deworming treatment. However, they were potentially affected by epidemiological spillovers generated by deworming treatment, since treating infected individuals means they are less likely to pass on worm larvae into the environment through fecal matter (the usual route of transmission for intestinal helminth infections). Ozier finds evidence that the youngest children (those under 2) gain substantially a full 10 years after deworming treatment in terms of their cognitive performance and academic test scores, with average gains of roughly half a school year of learning. This finding reinforces the

results in [Miguel and Kremer \(2004\)](#) about the potentially large magnitude of positive deworming treatment externalities in an area with high rates of worm infections; infection rates at baseline in this region of western Kenya were over 90%.

An implication of these externality effects is that research on infectious diseases—or other types of health or economic interventions—that does not account for externalities is likely to underestimate total program effects, both by potentially understating differences across the treatment and control groups (if the control group is gaining relative to the counterfactual of no exposure to spillovers), and by missing out on the spillovers entirely, thus doubly undercounting effects. The existence of treatment externalities thus makes cluster randomized designs—that treat most or all individuals in a given area, and consider the entire unit “treatment” in the analysis—more attractive than individually randomized designs in such settings, since treatment spillover benefits are at least partially “internalized” within the cluster (although as [Miguel and Kremer, 2004](#) show, some spillovers may extend beyond the cluster and these could be important to consider as well). We discuss this issue in greater detail below, but the presence of sizeable treatment externalities is a possible explanation for why several of the early studies of deworming treatment impacts on growth and cognition—all of which randomized across individuals within the same community or school—tended to find quite small (although typically positive) effects (see [Dickson et al., 2000](#)), namely, that the control group gained considerably from the intervention, dampening effects. In contrast, both of the large cluster randomized experiments on deworming discussed below (the [Miguel and Kremer, 2004](#) study, as well as the [Alderman et al., 2006b](#) project in Uganda) find both large short-run and long-run impacts of deworming on participant outcomes. The fact that many of the early deworming RCTs were conducted by nutritionists (rather than epidemiologists) might help explain the minimal attention paid to these issues in those studies.

A large number of studies within economics—including both health and nonhealth studies—have subsequently utilized the same basic empirical approach as [Miguel and Kremer \(2004\)](#) in order to estimate the magnitude of treatment externalities. Some of these studies modify the estimator to focus on the share of individuals within one’s social network that are affected by a treatment, rather than relying on geographic distance per se, as in the original analysis. In the health sector, this includes studies of mental health ([Baird et al., 2013](#)), water treatment ([Ziegelhöfer, 2012](#)), learning about HIV results ([Godlonton and Thornton, 2012](#)), community monitoring of health clinic performance ([Bjorkman and Svensson, 2009](#)), risky sexual behavior ([Dupas, 2011b](#)), child nutrition ([Zivin et al., 2009; Fitzsimons et al., 2012](#)), family planning ([Joshi and Schultz, 2013](#)), and malaria prevention ([Tarozzi et al., 2014; Dupas, 2014b](#)), as well as a study of take-up of the deworming treatments themselves within a social network ([Kremer and Miguel, 2007](#)), among many other related research studies.

As might be expected, given the diversity of health conditions and behaviors that have been examined using these methods, the magnitude and range of externalities vary

considerably across cases. However, it is worth mentioning the estimated effects in some of these cases. The case of malaria is particularly important, given how widespread the condition is in many low-income regions (especially in Africa) and its contribution to the total global burden of disease. Both of the malaria studies in economics mentioned above find suggestive evidence that positive externalities are generated when households use insecticide-treated bed nets, although neither has adequate statistical power to reach definitive conclusions (Tarozzi et al., 2014; Dupas, 2014b). In contrast to deworming, the malaria spillover benefits tends to be localized within a community, and it appears to within 20–30 m from the household using the net (Tarozzi et al., 2014). This information on the magnitude and geographic extent of spillovers can be important for both public health planners, as well as for those considering the desirability of large public subsidies for these, or other, health interventions.

As alluded to above, In other recent work economists have moved beyond studying epidemiological spillovers directly (as in the deworming and malaria cases), and have begun to explore spillovers through social networks in terms of technology adoption and behavioral change (as in Kremer and Miguel, 2007; Dupas, 2011b; for instance), and also the possibility that spillover could occur through channels other than epidemiology or social influence. For instance, one direct way that externalities might occur is through the sharing of medical treatment between those assigned to treatment and those assigned to control; in the case of a treatment such as iron supplementation which has limited side effects, this is something that might be considered quite low risk among participants (see the discussion of Thomas et al., 2003, 2006; Bobonis et al., 2006 for studies on iron supplementation in this literature).

Recent research has made methodological progress in understanding how to most efficiently estimate externality effects, and how to address the possibility of nonlinearities in the relationship and complementarities with local treatment decisions. Bhattacharya et al. (2013) exploit experimental variation combined with detailed geospatial information to estimate how the local subsidy rates faced by others affect insecticide-treated mosquito nets (ITN) use in Kenya, and show that there are important nonlinearities in the subsidy incidence. The issue of possible nonlinearities in social effects is raised as a possibility in both Miguel and Kremer (2004) and Kremer and Miguel (2007) but in neither study was there sufficient statistical power to reject linear specifications. Baird et al. (2014) discuss the optimal design of experiments to estimate spillover effects in settings where it is possible to randomize the intensity of treatment within clusters, and then randomly assign individual treatment conditional on this intensity. They include calculations of statistical power to detect externality effects given program parameters, which is useful for those prospectively designing experiments with this aim.

In areas beyond health, spillover effects and related general equilibrium effects are increasingly being studied in a wide range of sectors including in the study of cash transfer programs, microfinance programs, and beyond, demonstrating the analytical usefulness of

these approaches to economics research as a whole; [Muralidharan and Sundararaman \(2015\)](#) present an application to education, and [Angelucci and Di Maro \(2015\)](#) provide further discussion of such studies across subfields within development economics.

## 2.2 Experimenting to understand the determinants of health behavior: beware of measurement

Like other field experiments in development economics, experiments focusing on health topics have often relied on original data collection—including individual and household survey data, biomarker data, as well as data from clinics and schools—in the analysis. As they were with research design issues, these studies have also been highly innovative in their development of new data collection methodologies, as well as in clarifying some of the potential biases that could arise from these different types of original data collection. We discuss these different concerns—namely the direct effect of being surveyed on responses, social desirability bias, and the reliability of health self-reports—in turn in this subsection.

The simplest and potentially most pervasive form of bias from original data collection would occur if any act of being surveyed itself affected subsequent responses and, even more importantly, behaviors. [Zwane et al. \(2011\)](#) quantifies the possible extent of this bias using data from multiple data collection activities in development economics, all of which featured some randomized variation in the frequency with which different groups of households or individuals were surveyed followed by administrative data collection on the outcome of interest, and show that the experience of being surveyed can often affect subsequent behavior in health studies, as well as in microfinance projects.

In the context of the health data that was featured in their analysis, the authors show the randomly chosen individuals in rural Kenya communities who were surveyed more frequently regarding their children's health status (here, the diarrhea outcomes and other health dimensions for infants) were significantly more likely to change their behavior in the direction of making more investments in their children's health, specifically, in the use of a point-of-use chlorine disinfectant for household water. These behavioral responses were large in magnitude and statistically significant among the households surveyed at high frequency (biweekly) relative to those surveyed infrequently (every 6 months), with a near doubling in use of chlorine disinfectant. This response also appears to have led to large reductions in reported diarrhea, and they are large enough to change the estimated effect of an ongoing water investment campaign (namely, spring protection) in the same region. Taken together, the authors suggest that frequent surveys may serve as a reminder to households to engage in particular health practices, similar to the effect that has been documented for explicit reminders through mobile phone and other means (for instance, see [Pop-Eleches et al., 2011](#)).

This has extremely important implications for health studies. While many economics studies collecting original data utilize relatively infrequent data collection (presumably for

reasons of cost), some like those discussed in [Zwane et al. \(2011\)](#) do make use of high frequency data collection, and such approaches are actually the “standard” in many public health studies, such as those studying diarrhea outcomes in children (the health data in Zwane et al. was modeled on these approaches). Data reliability might be improved to the extent that data can be collected less frequently from a larger sample of individuals, or to the extent that more “passive” forms of data collection, such as from administrative records or “big data” sources (such as mobile phone usage), rather than high frequency enumerator visits. An alternative that is increasingly employed in field experiments in development economics is the creation of a “pure control” group of households or communities who are not contacted by the research team or surveyed until the very end of the study, when outcome measures are collected; for an example of a study that uses this approach, see [Muralidharan and Sundararaman \(2011\)](#). These individuals are thus unlikely to have been affected by the process of data collection, and any such bias on the “regular control” group can also be quantified in this way.

A related but distinct concern with original data collection relying on surveys is the possibility that respondents will provide answers that they think the enumerators want to hear, what is known as social desirability bias, or experimenter demand effects. These are widely discussed in laboratory data collection in experimental economics, and are increasingly recognized as a major concern in field experimental data collection settings.

In many settings where sensitive health information is collected, researchers are increasingly creating “private” situations within the data collection encounter for them to enter in such data in a way that cannot be immediately verified by the enumerator (for instance, see [Baird et al., 2008](#)). These concerns may be particularly pronounced when it comes to reproductive health and sexual health topics. To address these concerns, scholars have recently been quite creative in employing enumerators who are more likely to elicit truthful responses from respondents, e.g., [Robinson and Yeh \(2011, 2012\)](#) hire former sex workers to privately survey other sex workers on their sexual practices and decision-making.

An alternative approach that creates privacy for respondents is a survey technique called list randomization. List randomization, also known as the item count or unmatched count technique, allows respondents to report on potentially sensitive behavior without allowing the researcher or surveyors to identify individual responses. In practice, some proportion of survey respondents are randomly selected to receive a short list of statements (e.g., general health choices and outcomes, say) and asked to report how many, but not which, statements are true. The remainder of the survey respondents are presented with the same list of statements and one key additional statement designed to capture sensitive behavior (e.g., regarding sensitive sexual behavior). By subtracting the mean number of true statements in the first group from the mean number of true statements in the second group, researchers can estimate the proportion of the sample that engages in the sensitive behavior. This approach has been widely used to study health

behaviors in many contexts (see [Droitcour et al., 1991](#); [LaBrie and Earleywine, 2000](#); [Chong et al., 2013](#)), as well as sensitive behavioral choices in other spheres ([Karlan and Zinman, 2011](#)), though an obvious limitation of this method is that it only generates group-level outcomes, not individual-level ones.

Even when techniques such as creating a private space for survey respondents, employing more appropriate enumerators, and list randomizations are used, there remain important concerns about the validity of self-reported health behaviors, especially in sensitive areas such as sexual and reproductive health. A growing number of studies have documented a sharp divergence between self-reported sexual behavior and objectively measured infection status. In data collected from over 10,000 adolescents in Western Kenya, [Duflo et al. \(2015a\)](#) find that 4.6% of girls and 4.8% of boys who report that they never had sex test positive for Herpes Simplex Virus type 2 (HSV2), a sexually transmitted infection, suggesting pervasive misreporting. [Gong \(2015\)](#) uses field experimental data from Kenya and Tanzania in the context of an HIV/AIDS related testing and information campaign, and shows that self-reported sexual behavior becomes less risky for individuals who were informed that they had tested HIV-positive, even while their incidence of STI infections—a more reliable measure of risky sex—increases significantly.

### 2.3 Research transparency, registration, and preanalysis plans

There is growing awareness in economics fields that current research methods and practices can sometimes produce misleading bodies of evidence ([Miguel et al., 2014](#)), and in many ways this growing awareness in the social sciences parallels earlier trends in medical research, as we discuss below. For instance, there is growing evidence documenting the prevalence of publication bias in economics, as well specification searching, and widespread inability to replicate empirical findings ([Rosenthal, 1979](#); [Simonsohn et al., 2014](#)). While some of these issues have been widely discussed within the fora of economics for some time (see [Leamer \(1983\)](#), [Dewald et al. \(1986\)](#) and [DeLong and Lang \(1992\)](#), among others), there has been a recent flurry of activity documenting these problems, and also generating ideas for how to make research more transparent and reproducible.

A leading proposed solution to the problem of publication bias is the registration of empirical studies in public registry. This would ideally be a centralized database of all attempts to conduct research on a certain question, irrespective of the nature of the results, and such that even null (not significant) findings are not lost to the research community. The most high profile attempt at a registry within Economics, and indeed, across the social sciences, is the new AEA Randomized Trial Registry ([AEA RCT Registry, 2013](#)), which was launched in May 2013.

In another example of intellectual exchange across economics and health, the AEA registry was explicitly inspired by existing registries for medical trials. Clinical trials began being registered in large numbers in the 1990s, but the proportion registered has

increased dramatically since roughly 2005, when more stringent requirements were placed on medical researchers seeking to publish in leading medical journals, as well as by government medical regulatory authorities. While recent research in medicine finds that the registry has not eliminated all underreporting of null results or other forms of publication bias and specification searching (Laine et al., 2007; Mathieu et al., 2009), they do over time help constrain inappropriate practices, and at a minimum the existence of a registry allows the research community to quantify the extent of these problems. It also helps scholars locate studies that are delayed in publication, or are never published, helping to fill in gaps in the literature and thus resolving some of the problems of “disappearing” null results identified in Franco et al. (2014).

Though it is too soon after the adoption of the American Economic Association (AEA)’s trial registry for randomized controlled trials to measure the impact, the registry is being used by many empirical researchers: in its first 2 years, over 500 studies conducted in over 80 countries had been registered, and the number continues to rise each month. In addition to the AEA’s registry, several other registries have recently been created across the social sciences, although they have received fewer studies and less attention so far. These include registries created by the International Initiative for Impact Evaluation (3ie) for international development studies the Registry for International Development Impact Evaluations, (RIDIE), launched in September 2013, and the Experiments in Governance and Politics (EGAP) registry, also created in 2013.

Parallel to the trend in the preregistration of studies, support has grown for including preanalysis plans (PAP’s) that can be posted and time stamped even before data are collected or are otherwise available for analysis in prospective studies (Miguel et al., 2014). While there were scattered earlier cases of preanalysis plans being utilized in the social sciences (most notably Neumark, 2001), the numbers of published papers using prespecified analyses have grown rapidly in the past few years, mirroring the rise of studies on the AEA registry. Some of these early uses of preanalysis plans in Economics are in health economics, most notably the influential Oregon Health Insurance experiment studied in Finkelstein et al. (2012). This is not unexpected given how widespread preregistration of studies and analysis plans has become within medical research. However, most economics studies using preanalysis plans have been within development economics (see Casey et al., 2012; Olken et al., 2014 among others). Casey et al. (2012) show how the lack of a preanalysis plan might have provided sufficient latitude for an unscrupulous researcher to report a wide range of different—and erroneous—conclusions using the same data, heightening concern about the possible extent of specification searching and biased reporting even in studies using randomized experimental designs.

There remain many open questions about whether, when, and how preanalysis plans could and should be used in Economics research, with open debates about how useful they are in different subfields of the discipline (Olken, 2015; Coffman and Niederle, 2015). Yet even among these authors, who are critical about widespread adoption of

preanalysis plans in all cases, there appears to be a growing consensus that, in certain situations—such as large-scale randomized trials that are expensive or difficult to repeat, and/or cases where a government, policymaker, or corporation has a vested interest in the outcome—preanalysis plans can increase the credibility of reporting and analysis. To follow-up on [Leamer's \(1983\)](#) famous pun, preanalysis plans can help keep the “con” out of randomized controlled trials.

### **3. EXPERIMENTAL ESTIMATES OF THE IMPACT OF HEALTH ON INDIVIDUAL PRODUCTIVITY**

There is a growing literature within Economics that uses experimental variation to estimate the impact of health status on various measures of individual productivity. In this section, we focus on the experimental studies on this topic, and largely ignore the vast observational literature on these issues, a literature that spans economics, public health, epidemiology, and medical trials.

It is useful to divide the emerging experimental health economics literature on this topic into three groups: first, those studies that directly examine the impact of improved health status on current adult labor productivity and other economically relevant outcomes; second, those studies that examine the impact of improved child health and nutrition on current educational and other outcomes; and finally, those studies that estimate longer-term persistent effects of earlier health investments on later productivity measures and other life outcomes. We consider these in turn below.

#### **3.1 Impacts of adult health and nutrition on productivity**

An important early experiment that estimated the effect of adult health status on contemporaneous measures of productivity and individual well-being is [Thomas et al. \(2003, 2006\)](#), the Work and Iron Status Evaluation (WISE). The intervention aimed to address iron deficiency anemia (IDA), one of the world’s most widespread health and nutritional problems. IDA is well known to lead contribute to physical weakness and lower aerobic capacity, and thus could plausibly affect individual labor productivity. The WISE study features a randomized evaluation of iron supplementation (weekly supplements of 120 mg of iron) plus deworming to a large sample of over 17,000 adults in Indonesia, with ages ranging from 30 to 70 years old. Since roughly 30% of the sample were infected with intestinal helminths at baseline, the impact of the intervention should be interpreted as the combined effect of iron and deworming.

It is worth noting that WISE features an unusual study methodologically within economics—although not medical research—in that it was carried out as a double-blinded experiment, i.e., the control group received placebo pills of identical appearance. This might limit any behavioral responses to the treatment that are due to the fact that beneficiaries know they are receiving treatment. While this sort of design is considered

the ideal for medical trials, it is debatable whether it constitutes a similar “gold standard” for social science research studies, where endogenous behavioral responses are often central to the theoretical framework motivating a given study. Double-blinding is possible for a relatively simple intervention, such as the iron supplementation and deworming in [Thomas et al. \(2003, 2006\)](#), but it is also often logically infeasible in more complicated interventions, or those in which participants themselves are called upon to make decision (for instance, in the studies of technology take-up discussed below).

[Thomas et al. \(2003, 2006\)](#) follow-up participants in the WISE study for 6 months, and focus on the intention-to-treatment estimates of program impact. They first document that iron status does improve significantly in the treatment group, with particularly large gains among those whose baseline hemoglobin (Hb) level was particularly low (below 12 g/dL, a common cut-off for anemia). The heterogeneity in Hb gains as a function of baseline deficiency motivates an estimation strategy that is similar to a difference-in-difference-in-differences (triple difference) approach: outcomes are compared between the treatment and control groups, over time (posttreatment versus baseline), across groups that had relatively low Hb at baseline (below 12.5 g/dL) versus relatively high Hb. This approach provides more statistical power than the simple treatment versus control difference, since a large share of individuals in the treatment group, namely those with relatively high Hb at baseline, do not experience any gains in Hb as a result of the intervention, and thus would not necessarily be expected to experience any gains in productivity.

In their 2006 working paper ([Thomas et al., 2006](#)), the authors report evidence of sizeable and statistically significant gains in a range of economic and wellbeing outcomes, with effects particularly large for males (although they are generally of the same sign for females, although smaller in magnitude). The probability that individuals are not working falls significantly by between 3 and 5 percentage points for both males and females, there is suggestive evidence that total earnings increase for males, and statistically significant gains in self-employed total earnings and hourly earnings (which is similar to a wage measure) for males, as well. There are also substantial gains in psycho-social outcomes for both genders, with males finding less difficulty sleeping and having more energy and leisure time, and females feeling less anxious. Given the relatively low cost of iron supplementation and deworming, and authors argue that this investment could have a high economic return. Given these returns, a question remains why individuals are not already making these sorts of investments in iron, deworming or improved nutrition more broadly in the absence of the intervention. It is also worth noting that even where there is an intervention, take-up of the nutritional supplements can remain limited (see the discussion on the demand for double-fortified salt in India in [Section 5.3](#)). Thus the rigorous research design, large sample size, and rich set of outcome measures make the preliminary evidence of the WISE study some of the most provocative estimates of

the causal impact of improved adult health on contemporaneous economic productivity to date. As with all experiments the external validity of the results is unknown, and we hope that additional studies of this kind will be conducted in different settings to help further our knowledge on this important topic.

Iron deficiency anemia is a pervasive but rarely fatal health condition. [Thirumurthy et al. \(2008\)](#) usefully consider a much more severe disease, HIV/AIDS, and estimate impacts of the introduction of antiretroviral (ARV) treatment on individual labor productivity measures in a Kenyan site. This study is not an RCT, but given the paucity of evidence on the productivity impacts of contemporaneous health in developing countries, it is worth briefly describing. The introduction of ARVs at the individual level is determined by a “cut-off” value of the individual CD4 count (which captures how compromised the individual immune system is), and thus the study’s design exploits experimental variation, although for ethical reasons treatment was not provided in a randomized fashion across individuals. Incorporation into ARV treatment during this period (when ARV treatment was still rare in Kenya) was typically life-saving and thus any labor earnings can plausibly be considered a treatment effect relative to the counterfactual. The authors more conservatively compare earnings after treatment to those immediately before treatment, which is arguably a lower bound on true effects. Due to the relatively high cost of treatment and limited number of individuals incorporated into the sample over the study period, they compare 266 households with at least one HIV positive individual to 503 other households representative of the local population.

[Thirumurthy et al. \(2008\)](#) estimate large and statistically significant gains of over 50% in individual labor supply up to 6 months after the start of ARV treatment, with large impacts on total earnings. There are also important within-household effects, as the labor supply of other individuals in the households, including children, fall after an adult begins treatment, suggesting that adult health status has important externalities for others in the household and may affect the human capital accumulation of the next generation. While perhaps not surprising given that most of the treated individuals would have passed away in the absence of ARV treatment, this study provides further evidence on the important economic consequences of large health shocks.<sup>1</sup>

An important share of the morbidity burden among adults in poor countries, especially Sub-Saharan Africa, comes from malaria. Yet there is little rigorous evidence to date on the productivity impact of malaria. To the best of our knowledge, two field experiments were set-up in the last decade to estimate those impacts. In what follows we describe their findings but also the challenges they have met. Indeed, the dearth of evidence on this issue stems primarily from the inherent difficulty in measuring productivity

<sup>1</sup> [Fox et al. \(2004\)](#) provide related nonexperimental evidence on the labor productivity effects of HIV/AIDS, in their case on a Kenyan tea plantation.

precisely without very large samples, and from the difficulty in changing health but nothing else at the same time, i.e., of isolating the health channel.

One of the two field experiments we are aware of is [Fink and Masiye \(2015\)](#), who randomized access to bed nets among cotton-growing households in a rural area of Zambia with highly endemic malaria. They compare a control arm to an arm where households could obtain additional bed nets for free, and to an arm where they could get bed nets at subsidized prices (through an agricultural loan program). On average, 2.4 nets were distributed in the free distribution group and 0.9 in the net loan group. As found previously in other contexts ([Dupas, 2009](#)), 90% of farmers used their nets at follow-up in both intervention arms, and as a result the interventions led to a large drop in self-reported all-cause morbidity (by 40–42%) and the odds of self-reported confirmed malaria by 53–60% ([Fink and Masiye, 2012](#)).

A central methodological issue is that measuring the productivity of farming households in low income settings is challenging. [Fink and Masiye \(2015\)](#) rely on self-reported yield and find an increase in total farm output of 14.7%, suggesting large positive impacts of reducing malaria cases. However, computing yields is nontrivial for farm households, since assessing the value of unpaid labor inputs across individuals within the household is difficult, and there may also be a concern that health gains might have influenced the accuracy of reporting, even beyond any actual gains in yields.

These estimated gains are obtained conditional on the baseline yield, which unfortunately appears quite imbalanced across arms, possibly owing to the relatively small sample size. The randomization was clustered at the farming inputs distributor level, so a cluster corresponded to 11 randomly selected farmers from among a given distributor's clients, with only 49 clusters in total, 15 assigned to control, 15 to the free net arm, and the remainder (19) to the loan program. At baseline, farms in the free net and net loan arms were, on average, larger and more productive than farms in the control group, despite relative balance with respect to household head characteristics. Given this, it is not clear that controlling for baseline values in yield—a de facto difference-in-differences design—is sufficient to recover the causal impact of the bed net on productivity, since it remains possible that these baseline differences also translate into differential trends over time, reflect other unobservables, or capture differences in how susceptible the households are to rainfall shocks. The existence of reported yield data over only a single season is also potentially problematic if climatic conditions in that year might have disproportionately favored particular types of farmers, and thus particular treatment arms (given the baseline imbalance), making it difficult to even argue that effects are likely to be lower or upper bounds. Thus while the analysis suggests large positive impacts of treating malaria on productivity, there are reasons to be cautious in our interpretation of the findings.

The second field experiment attempting to estimate the productivity impacts of malaria was conducted by [Dillon et al. \(2014\)](#) with sugar cane cutters in Nigeria. Here the difficulty in observing productivity is less of an issue, since sugar cane cutters

are paid a piece rate for every measured “rod” of cane cut, where a “rod” (approximately 2 m in length) is a physical standard that a worker’s supervisor carries to the final dropping point once completed by the worker. At the end of each day, the worker’s output for that day is entered on his personal ‘blue card’ and is signed off on by both the supervisor and worker. The plantation thus keeps records of the daily output (quantity cut), the days worked, and the total earnings for each worker, which means that getting access to this data provides the researchers with a high-quality and objective measure of productivity.

The key constraint facing the research team was instead in the type of health intervention that could feasibly be carried out. It seems that the ideal research design, randomizing access to malaria prevention over the entire period across workers, was not possible. Instead, in the experiment, 800 workers who had been hired for a 6-week harvest season by a large sugar cane plantation were called in for a “medical visit” at some point over the 6-week period, with the exact date on which the visit happened randomized across workers. During the medical visit, workers were tested for malaria and, if positive, treated with highly effective antimalarial treatment. On average around 30% of workers were found positive and subsequently treated. Ideally, there would be a control group of individuals also positive but not treated, and comparing those to the treated would provide an estimate of the impact of malaria detection and treatment on labor productivity. Unfortunately, the experimental design did not generate that, as those not sampled for the medical visit were simply not tested at all, making it impossible to assess their malaria status.

The best the researchers can do here is to compare workers who were tested early on in the study period to those tested later on. Doing so, they find a large intention-to-treat effect: those sampled for an early test are 15% more productive on the plantation during the three weeks following the medical visit compared to those tested later on. This is a large effect, and remarkably similar to the nearly 15% increase in farm productivity after bed net distribution reported in the Zambia study discussed above ([Fink and Masiye, 2015](#)). (There are some possible discrepancies, though, since the larger measured malaria reductions in the Zambia data might also imply larger productivity impacts, to the extent that the farm labor tasks in both cases are largely comparable.) Moreover, the intervention evaluated in the Nigerian case is somewhat peculiar: it provides each worker a one-time testing and treatment opportunity on a preassigned day regardless of how they feel on that day. This is quite different from most real-world malaria policies one might imagine. In particular, this approach will likely have a smaller productivity impact than making malaria testing and treatment free and easily available for workers on any day that they feel ill, suggesting that the 15% effect in [Dillon et al. \(2014\)](#) is a lower bound of providing treatment on a permanent basis. That said, the intervention studied may have the advantage of catching and resolving some cases that affect productivity but are not sufficiently severe to lead workers to seek treatment on their own.

Other recent studies have exploited empirical settings outside of the family farm where individual worker productivity can be accurately measured. One of the most noteworthy recent contributions is [Adhvaryu et al. \(2014a\)](#). This study uses data from garment-factory workers in India, whose productivity is accurately measured by managers in the natural course of running the assembly line. The health issue they focus on is exposure to fine particulate matter (PM2.5) in the factory, which is monitored at high frequency by multiple sensors on the factory floor. Worker rotation to different tasks within the factory leads to differential exposure, as does natural variation in pollution levels. Exposure to fine particulate matter leads to a range of respiratory and cardiovascular problems in the short-run, including acute constriction of blood vessels, and long-run exposure is linked with severe health and mortality risks.

The central finding is that higher levels of PM2.5 exposure significantly reduce factory worker productivity: an increase of one standard deviation increase in PM2.5 air pollution (roughly  $45 \mu\text{g}/\text{m}^3$ ) reduces productivity by 6%. This provides further evidence that higher morbidity—in this case, along a health dimension different from other existing studies—is associated with lower earnings. The authors also report that workers working under experienced factory managers also appear to suffer less from high pollution levels, although the precise mechanisms underlying this pattern are unclear. Air pollution is an unfortunate fact of life for hundreds of millions in the rapidly growing cities of Asia and Latin America, and increasingly in Africa, and one that appears to only become more severe over time, suggesting that these findings have broad scientific import and policy relevance.<sup>2</sup>

### **3.2 Impacts of child health and nutrition on education**

A distinct subliterature estimates effects of child health and nutritional investments on contemporaneous educational outcomes; [Glewwe and Miguel \(2008\)](#) provide a thorough review of both the experimental and nonexperimental research in this area. Here we focus on a selection of the experimental studies in this area. This is actually a vast literature that crosses many disciplines, and it is beyond the scope of this survey to cover all relevant studies. We focus mainly recent studies within economics, but also discuss some related contributions from other field.

Many of the earliest randomized studies by nutritionists and other public health researchers focused on the impacts of specific nutrients that were lacking in children's diets. Studies in India and Indonesia by [Soemantri et al. \(1989\)](#), [Soewondo et al. \(1989\)](#), and [Seshadri and Gopaldas \(1989\)](#) found large and statistically significant impacts

<sup>2</sup> In a companion paper, [Adhvaryu et al. \(2014b\)](#) estimate large negative effects of higher temperatures on garment-factory worker productivity, a finding with potential importance for the possible economic productivity impacts of future climate change and global warming.

on cognitive development and school performance of iron supplementation among anemic children, but a study by [Pollitt et al. \(1989\)](#) found no such impact in Thailand. See [Nokes et al. \(1998\)](#) for a more complete survey of the related iron supplementation literature.

Other early studies focused on parasitic infections, especially intestinal parasites. [Kvalsig et al. \(1991\)](#) examined whipworms and other parasites in South Africa and found that drug treatments had some effect on cognitive and education outcomes, but some impacts were not statistically significant. [Nokes et al. \(1992\)](#) evaluated treatment for whipworms in Jamaica and concluded that some cognitive functions improved from the drug treatment, but others, particularly those related to academic performance in schools, appeared not to have changed substantially. Overall, the early experimental literature on the impact of treatment for intestinal parasites on child growth and cognition did not reach strong conclusions, as argued in the [Dickson et al. \(2000\)](#) survey and in the more recent Cochrane review on the topic ([Taylor-Robinson et al., 2012](#)). One possible reason why many of the early experimental deworming studies show limited impacts is that they carried out randomized treatment within school communities, creating the possibility that positive treatment externalities experienced by children in the control group lead to attenuated treatment effects, as discussed earlier and in [Miguel and Kremer \(2004\)](#). Many of these studies also have relatively small sample sizes, such as 210 children in the South African study and 103 in the Jamaican study. Other experimental studies (not reviewed here) include education interventions combined with health interventions, so the impact of the health intervention by itself cannot be credibly assessed.

Other studies have focused on general food supplementation to supply calories and protein. The most well-known of these is the INCAP study ([Pollitt et al., 1993](#); [Martorell et al., 1995](#)) initiated in four Guatemalan villages in 1969, two of which were randomly selected to receive a porridge (atole) high in calories and protein while the other two villages received a drink (fresco) with less calories and no protein. Follow-up studies over the next three decades appear to show sizeable effects on later cognitive outcomes from providing the atole to mothers and young children, and we discuss these in greater detail below.

A number of recent randomized experiments have also been carried out by economists on the impact of health interventions on educational outcomes. These studies also typically evaluate actual interventions carried out by real-world nongovernmental organizations (NGOs) or governments, and as such the findings of these studies may be of particular interest to policymakers in less developed countries. These are in contrast to several of the studies discussed above, which were often small-scale researcher implemented interventions. Many of these evaluate school-based health or nutrition interventions which some have argued may be among the most cost-effective approaches for delivering health and nutrition services to children in less developed countries ([Bundy and Guyatt, 1996](#)).

[Miguel and Kremer \(2004\)](#), discussed at length in [Section 2.1](#) evaluates the impact of school-based mass treatment for intestinal worms using inexpensive deworming drugs. The study is based on a sample of 75 primary schools with a total enrollment of nearly 30,000 children, a much larger sample size than most other studies in this literature. The sampled schools were drawn from areas where there is a high prevalence of intestinal parasites among children. Worm infections—including those caused by hookworm, roundworm, whipworm as well as schistosomiasis—are among the most widespread diseases in less developed countries: recent studies estimate that 1.3 billion people worldwide are infected with roundworm, 1.3 billion with hookworm, 900 million with whipworm, and 200 million with schistosomiasis. Infection rates are particularly high in Sub-Saharan Africa ([Bundy et al., 1998](#); [World Health Organization, 1993](#)). Geohelminths—hookworm, roundworm, and whipworm—are transmitted through poor sanitation and hygiene, while schistosomiasis is acquired by bathing in infected freshwater. School-aged children typically exhibit the greatest prevalence of infection and the highest infection intensity, as well as the highest disease burden, since morbidity is related to infection intensity ([Bundy, 1988](#)).

The educational impacts of deworming are considered a key issue in assessing whether the poorest countries should accord priority to deworming, but until recently research on these impacts has been inconclusive (see [Dickson et al., 2000](#) for a survey). Indeed, earlier randomized evaluations on worms and education suffer from several important methodological shortcomings that may partially explain their weak results. Earlier studies randomized the provision of deworming treatment within schools to treatment and placebo groups, and then examine the impact of deworming on cognitive outcomes. However, the difference in educational outcomes between the treatment and placebo groups understates the actual impact of deworming if placebo group pupils also experience health gains due to local treatment externalities (due to breaking the disease transmission cycle). The earlier studies also failed to adequately address sample attrition, an important issue to the extent that deworming increases school enrollment.

The study by Miguel and Kremer finds that absenteeism in treatment schools was 25% (7 percentage points) lower than in comparison schools and that deworming increased schooling by 0.14 years per pupil treated (on average). This is a large effect given the low cost of deworming medicine; the study estimates an average cost of only US\$ 3.50 per additional year of school participation. The finding on absenteeism does not reflect increased school attendance on the part of children who attend school only to receive deworming drugs, since drugs were provided at only two preannounced days per year, and attendance on those two days is not counted in the attendance analysis. There is no statistically significant difference in treatment effects between female and male students.

Somewhat surprisingly, despite the reduction in absence, no significant impacts were found in relation to student performance on academic tests. It is unclear what exactly is

causing this discrepancy, although one possibility is that the program led to more crowded classrooms and that this may have partially offset positive effects of deworming on learning in the treatment schools.

The schooling data in [Miguel and Kremer \(2004\)](#) are noteworthy from a measurement perspective. School attendance was collected at sample schools by survey enumerators on unannounced days four to five times per year, rather than relying on school registers (which are often thought to be unreliable) or on parent reports in household surveys, as done in most of the previous literature. Efforts were also made to follow children who transferred to other schools in the same Kenyan district. This yields a more detailed and reliable measure of school participation than the data available from most other studies. [Bobonis et al. \(2006\)](#) and [Vermeersch and Kremer \(2004\)](#) use similar measures of school attendance.

The authors found that child health and school participation—i.e., attendance, where dropouts are considered to have an attendance rate of zero—improved not only for treated students but also for untreated students at treatment schools (roughly a quarter of pupils in treatment schools chose not to receive the deworming medicine) and for students at nearby primary schools located within 3 km. The impacts on neighboring schools appear to be due to reduced disease transmission brought about by the intervention, an epidemiological externality. Econometric identification of the cross-school treatment spillovers on the worm infection rate relies on the randomized design of the project: conditional on the total local density of primary school pupils, there is random exogenous variation in the number of local pupils assigned to deworming treatment through the program. A key finding of the paper is that failure to take these externalities (or spillovers) into account would lead to substantial underestimation of the benefits of the intervention and the cost effectiveness of deworming treatment.

[Miguel and Kremer \(2014\)](#) document a coding error in the construction of the variables used to measure treatment externalities at distances between 3 and 6 km from each school; correcting this issue weakens the statistical significant of externality effects on worm infections at this distance but does not affect other results in the original paper.<sup>3</sup>

[Bobonis et al. \(2006\)](#) conducted a randomized evaluation in India of a health program that provided iron supplementation and deworming medicine to preschool children age 2–6 years in 200 preschools in poor urban areas of Delhi. Even though only 30% of the sampled children were found to have worm infections, 69% of children had moderate to severe anemia according to international standards. After 5 months of treatment, the authors found large weight gains and a reduction of one-fifth in absenteeism, a treatment effect similar to the estimated school participation effect in the [Miguel and Kremer \(2004\)](#)

<sup>3</sup> For further discussion of the implications of the updated [Miguel and Kremer \(2004\)](#) findings, refer to [Aiken et al. \(2015\)](#), [Davey et al. \(2015\)](#), and [Hicks et al. \(2015\)](#).

study in Kenyan primary schools. The authors attempted to obtain estimates after 2 years, but high sample attrition and apparently nonrandom enrollment of new children into the preschools complicated attempts to obtain unbiased longer term impact estimates. It also does not present data on any type of child learning, and thus is limited to examining anthropometric outcomes and school enrollment and attendance. Finally, because all children received a combined treatment of iron supplements and deworming medicine, the India study cannot distinguish between the separate impacts of these two treatments, similar to the [Thomas et al. \(2003, 2006\)](#) studies discussed above.

An arguably cleaner test is provided by [Chong et al. \(2016\)](#), who randomized access to materials promoting iron supplementation among a small sample of 219 Peruvian secondary schoolchildren, in an area where local stores had also been stocked with iron pills. Baseline rates of anemia were fairly high in this population, at over 40%. The authors find that the informational encouragement design was effective, such that the treatment group consumed significantly more iron over the 3-month intervention period; this in itself is an interesting finding, given the many challenges in changing health behaviors that we detail in this chapter. More importantly, the schooling performance of the sample is then tracked using administrative records, and the data indicate that among anemic children in the iron supplementation treatment group, academic test scores rose by a sizeable 0.4 standard deviation units, and that grade progression and aspirations both also improved.

Other recent studies paint a more mixed picture of the impacts of child health and nutritional interventions on educational outcomes. [Clark et al. \(2008\)](#) examine the impact of intermittent preventive treatment (IPT) of malaria among schoolchildren in a region of western Kenya with high perennial transmission of the disease. The treatment occurred roughly every 4 months with a combination of sulfadoxine-pyrimethamine and amodiaquine, and outcomes were compared to a placebo group over the course of a year. This approach is reminiscent of the malaria screening and treatment approach in [Dillon et al. \(2014\)](#), described above, in that it does not depend on individuals seeking out treatment during periods when they are already ill. Randomization occurred among roughly 5000 children aged 5–18 within 30 sample primary schools.

The study produced evidence of large positive impacts of IPT on anemia, as well as on some cognitive outcomes after one year. In particular, there are significant gains in two class-based tests of sustained attention. There is no evidence of impacts on hyperactive-compulsive behavior or on educational achievement. There are two issues worth keeping in mind when interpreting the results. The first is the relatively high rate of attrition over the course of one year, at 27%, although rates appear balanced between treatment arms. A second is the fact that there was simultaneously mass treatment for intestinal helminth infections in the sample schools, which the studies mentioned above have shown lead to school participation gains in a nearby area of western Kenya. Thus the treatment effects in the [Clarke et al. \(2008\)](#) study are conditional on this deworming regime, and it remains

unclear if anemia and educational gains would have been larger or smaller in magnitude in the absence of deworming. The possibility that treatments for infectious diseases might serve as complements (or substitutes) is important from a public policy standpoint, but little is known about these interactions in practice; this remains a promising area for future research.

The expansion of large-scale national social welfare programs, often linked to conditional cash transfers, in many Latin American and other developing countries has provided opportunities to explore alternative ways to improve early child development outcomes. [Amarante et al. \(2016\)](#) exploit an eligibility discontinuity in the design of a cash transfer program in Uruguay to show that receipt of a large transfer leads to substantial reductions in the incidence of low birth weight among the children of beneficiaries. Beyond the cash itself, the contact with recipients induced by the program allows for opportunities to deliver additional interventions. [Attanasio et al. \(2014\)](#) study one such effort in Colombia as part of its *Familias en Acción* program, which carried out large-scale micronutrient supplementation as well as psycho-social stimulation of children aged 12–24 months, and assessed nutritional, health, and cognitive impacts. The stimulation consisted of weekly home visits, which are relatively expensive, as well as biweekly supplementation with iron, zinc, vitamin A, folic acid, and vitamin C, all carried out over the course of the 18 month study. The research consisted of a  $2 \times 2$  factorial design.

[Attanasio et al. \(2014\)](#) find that the nutritional supplementation alone did not have any impacts on measured child height, weight, hemoglobin, cognitive scores or receptive language, nor did it have any additional benefit in combination with stimulation. Psycho-social stimulation, on the other hand, did lead to significant improvements in the cognitive and language outcome measures, echoing similar findings from earlier work in Jamaica ([Gertler et al., 2014](#)). It remains possible that micronutrient supplementation would have larger impacts in poorer populations; recall that Colombia is a middle-income country and the study sample consisted of households receiving a generous cash transfer. Further research is needed to shed light on the generality of this finding.

### 3.3 Impacts of child health and nutrition on later outcomes

A third group of studies estimates long-run impacts of child health interventions on life outcomes, where again we focus on experimental studies in development economics.

We first examine a growing number of studies estimating long-run impacts of deworming. New evidence is rapidly accumulating on the positive long-run educational and socio-economic impacts of child deworming. A key lesson of [Miguel and Kremer \(2004\)](#) is that traditional individual-level randomized designs will miss any spillover benefits of deworming treatment, and this could contaminate estimated treatment effects. Thus cluster randomized designs provide better evidence. Three new working papers

with such cluster randomized designs estimate long-run impacts of child deworming up to 10 years after treatment; these effects on long-run life outcomes are arguably of greatest interest to public policymakers.

A main puzzle with the [Miguel and Kremer \(2004\)](#) Kenya deworming study is that increased school participation (primarily attendance, but also reduced dropping out) is not reflected in students' academic test scores or cognitive test scores. The authors present some cost-benefit analyses at the end of the paper that suggest that the intervention is cost-effective, but it is unclear exactly how to interpret these if the intervention does not increase learning of basic skills.

This issue is addressed in the follow-up study, [Baird et al. \(2015\)](#), which collects information on a wide range of adult life outcomes. [Baird et al. \(2015\)](#) followed up the Kenya deworming beneficiaries from the [Miguel and Kremer \(2004\)](#) study during 2007–09 and find large improvements in their labor market outcomes. The paper employs a conceptual framework building on the seminal health human capital model of [Grossman \(1972\)](#), which interprets health care as an investment that increases future endowments of healthy time. [Bleakley \(2010b\)](#) further develops this theory, arguing that how the additional time is allocated will depend on how health improvements affect relative productivity in education and in labor. [Pitt et al. \(2012\)](#) further note that time allocation will also depend on how the labor market values increased human capital and improved raw labor capacity, and that this in turn may vary with gender. They present evidence consistent with a model in which exogenous health gains in low-income economies tend to reinforce men's comparative advantage in occupations requiring raw labor, while leading women to obtain more education and move into more skill-intensive occupations.

Consistent with [Pitt et al. \(2012\)](#), the Kenya deworming program increased education among women and labor supply among men, with accompanying shifts in labor market specialization. Ten years after deworming treatment, women who were eligible as girls 25% more likely to have attended secondary school, halving the gender gap. They reallocate time from traditional agriculture into cash crops and entrepreneurship. Men who were eligible as boys stay enrolled for more years of primary school, work 17% more hours each week, spend more time in entrepreneurship, are more likely to hold manufacturing jobs, and miss one fewer meal per week. Since deworming treatment is inexpensive (at less than US\$ 1 per person per year), the authors estimate a large annualized financial internal rate of return of at least 32.2%. Many studies argue that early childhood health gains in utero or before age 3 have the largest impacts (for instance, [Almond and Currie, 2010](#)) and some have argued that health interventions outside a narrow biological window of child development will not have major effects. This evidence suggests that health interventions among school-aged children, which are too late in life to affect cognition or height, can have long-run impacts on labor market outcomes by affecting the amount of time people spend in school or work.

There are several noteworthy methodological features of the [Baird et al. \(2015\)](#) article. First, it remains unusual for studies to combine experimental designs and long-run 10 year follow-up longitudinal data, and in this case most individuals in the sample were between 19 and 26 years old at the follow-up. Second, the rate of attrition was quite low in the follow-up Kenya Life Panel Survey (KLPS). KLPS tracked a representative sample of approximately 7500 respondents who were enrolled in grades 2–7 in the Kenya deworming schools at baseline. Survey enumerators traveled throughout Kenya and Uganda to interview those who had moved out of local areas. The effective survey tracking rate in KLPS overall is 82.7%, and 84% among those still alive. These are high tracking rates for any age group over a decade, and especially for a mobile group of adolescents and young adults. Tracking rates are nearly identical and not significantly different in the treatment and control groups.

While the primary school children in the [Miguel and Kremer \(2004\)](#) sample were probably too old for deworming to have major impacts on brain development (and there was no evidence of such impacts), [Ozier \(2014\)](#) estimates cognitive gains 10 years later among children who were 0–2 years old when the deworming program was launched and who lived in the catchment area of a treatment school. These children were not directly treated themselves but could have benefited from the positive within-community externalities generated by mass school-based deworming. [Ozier \(2014\)](#) estimates average test score gains of 0.3 standard deviation units, which is equivalent to roughly half a year of schooling. This provides further strong evidence for the existence of large, positive, and statistically significant deworming externality benefits within the communities that received mass treatment.

[Croke \(2014\)](#) finds positive long-run educational effects of a program that dewormed a large sample of 1–7 years olds in Uganda, with statistically significant average test score gains of 0.2–0.4 standard deviation units on literacy and numeracy 7–8 years later. These are similar to the effect magnitudes estimated by [Ozier \(2014\)](#) in Kenya. The Ugandan program is one of the few studies to employ a cluster randomized design, and earlier evaluations of the program had found large short-run impacts on child weight ([Alderman et al., 2006a,b; Alderman, 2007](#)).

The long-run impacts of the well-known INCAP experiment in Guatemala are described in [Hodinott et al. \(2008\)](#), [Maluccio et al. \(2009\)](#), and [Behrman et al. \(2009\)](#). As mentioned above, INCAP provided substantial nutritional supplementation to two villages while two others served as a control, and the authors find evidence of very large and statistically significant gains in male wages by one-third, improved cognitive skills among both men and women, and even positive intergenerational effects on the nutrition of beneficiaries' children up to 35 years after the original project. This is a highly unusual and exceptional data collection effort, and it provides further evidence that childhood health and nutrition gains can have large returns in terms of adult labor productivity.

The pioneering INCAP study also has some limitations. In one sense, it has a sample size of only four villages since the intervention did not vary within villages, and it is unclear if all the existing studies fully account for the intraclass correlation of respondent outcomes in their statistical analyses, thus perhaps leading them to overstate the statistical significance of their findings. Second, strictly speaking, the control group also received an intervention, the fresco drink, albeit one with a relatively small benefit compared with what was received in the treatment group. Third, within each village receipt of the atole or fresco was voluntary, which implies that those who were treated were not a random sample of the population within each village. This means that the most convincing estimation strategy may be an intention to treat analysis, rather than direct estimation of the effect of child health on education. Finally, sample attrition is a major concern in both the 1988–89 follow-up and the most recent surveys, as more than one quarter of the original sample were apparently lost by 1988–89 and roughly 40% were lost by the time of the 35-year follow-up survey.

#### **4. ENVIRONMENTAL/INFRASTRUCTURAL DETERMINANTS OF HEALTH**

The [Adhvaryu et al. \(2014a\)](#) study on the impact of indoor air pollution on Indian factory worker productivity discussed above highlights the links between environmental conditions, health, and incomes, and the importance of environmental issues potentially extends far beyond their setting. In the United States and other countries, public health measures such as improved sanitation, provision of clean drinking water, and hookworm and malaria eradication campaigns have been shown to have played a key role in the massive improvements in child health and decreases in mortality observed during the twentieth century ([Cutler and Miller, 2005b](#); [Currie, 2000](#); [Bleakley, 2007, 2010a,b](#); [Galiani et al., 2005](#); and see [Deaton, 2013](#) for a review). The implementation of many such “infrastructural” public health investments is not only costly (at times prohibitively so, especially in low population density areas such as much of rural Africa), it also likely requires a minimal level of state capacity that may still elude many low-income countries. In this section, we discuss the relatively limited number of field experiments in developing countries that have recently studied the impact of environmental investments on health outcomes.

[Kremer et al. \(2011a\)](#) run an experiment to estimate the health consequences of improving water quality at the source in rural Kenya. Naturally occurring springs are an important source of drinking water in the area of their study (Western Kenya), yet few of them are “protected”—i.e., the water seeps from the ground, which implies that the water that pools is often contaminated with runoff from the surrounding area (often containing human and animal fecal matter and other sources of pollution). In contrast, a spring is considered protected if its source is sealed off and encased in concrete

so that water flows out from a pipe, where it can be collected by consumers before the water touches the ground.

[Kremer et al. \(2011a\)](#) exploit the randomized protection of 100 springs out of 200 in order to estimate the impacts of improved water quality on child health outcomes. They find that spring protection dramatically improves water quality: through follow-up visits conducted among a representative sample of households that had been identified as regular spring users at baseline, the authors find that fecal contamination of home-stored water was one quarter lower for those in the catchment area of a protected spring compared to those in the catchment area of an unprotected spring. As a result child health improves, with a 25% decrease in child diarrhea attributable to the spring protection intervention. Note that the study was not statistically powered to detect mortality impacts. The authors also tested for, but did not find any evidence of, epidemiological spillovers (in terms of water contamination) within 3 km of the protected springs.

The study goes on to study the demand for cleaner water, using a travel cost methodology developed in environmental economics to value nonmarket amenities. Specifically, [Kremer et al. \(2011a\)](#) estimate how much the usage of protected springs increases (relative to other water sources) after the intervention; large increases in usage, and especially a willingness to walk longer distances to reach such sources, would be consistent with increased demand for these sources. About three quarters of water trips were collected at the reference source (which was randomized into spring protection or control) at baseline, so while limited, there is still scope for some switching, and it is from this switching (or lack thereof) that the authors can back out a “revealed preference” estimate of willingness to pay for clean water. Using a discrete choice multinomial logit framework, the authors document that use of protected springs does increase significantly, but that magnitudes are not particularly large. Making standard assumptions on the value of water collector time, they estimate that the revealed preference willingness to pay for improved water quality is relatively low, at only US\$ 2.96 per year.

The authors then go on and translate this willingness to pay for water quality into a valuation for health. This exercise is challenging because the value placed on water quality may differ based on what the water is used for. While households may be willing to walk extra distance for clean drinking water, they may not do so for laundry water. [Kremer et al. \(2011a\)](#) do not have data on which water trip is for drinking water, thus they assume that the benefit from an additional trip to the clean water source is constant. In other words, they assume households trade off the cost of every water trip against child health benefits, potentially not only those made for drinking water. This is a relatively strong assumption since, as mentioned above, households in their sample conduct most of their water trips to the reference source to start with, and it

is unclear how much of the water they fetch at other (unclean) sources is actually for drinking.<sup>4</sup> Under this assumption, they estimate that willingness to pay is only US\$ 0.89 to avert one child diarrhea episode. Even stronger assumptions allow them to translate the diarrhea estimate into an implied mean value of averting one child statistical death of only US\$ 769, far lower than is commonly assumed by health policymakers, although this is plausibly an underestimate (given the discussion above). In [Section 5](#), we discuss the broader literature on the demand for health in developing countries, in which we discuss alternative and more direct approaches to estimate the demand for health products.

Besides access to clean water, access to sanitation (in particular connection to sewerage) has been shown to be an important determinant of health ([Cutler and Miller, 2005a](#)). Yet the type of sanitation possible in low-density settings does typically not involve sewage, and its potential for health impacts is unclear. In rural Orissa, in India, [Clasen et al. \(2014\)](#) estimate the impact of latrine construction and find only modest impact on latrine usage and no impacts on health. Similar results were found in a comparable experiment in rural Madhya Pradesh, also in India ([Patil et al., 2014](#)).

[Duflo et al. \(2015b\)](#) argue that these null results, rather than reflecting a low demand for health, may be due to the fact that sanitation by itself is not enough, and that water access may be a necessary complement. Indeed, maintaining a clean latrine without access to sufficient water quantity is difficult, likely reducing latrine usage. Moreover, given epidemiological externalities (such as those discussed earlier in the chapter), the private returns to sanitation may be limited if the majority of neighbors do not also have access to sanitation. This would mean that a sanitation program can only have a meaningful impact if combined with access to water *and* implemented at large scale (i.e., taken up nearly universally). [Duflo et al. \(2015b\)](#) estimate the effect of such a scheme, which provided household-level water connections, latrines, and bathing facilities to all households at once in approximately 100 Indian villages. The program was not randomized, but a before-after comparison suggests large impacts on hygiene and reductions in diarrhea rates.

In a cluster-randomized trial in rural Bangladesh, [Guiteras et al. \(2015\)](#) find that providing subsidies to the majority of the landless poor households increases ownership and usage of latrines among both subsidized households and their neighbors, which is also consistent with the presence of important epidemiological externalities connecting individual investment and usage decisions with those of neighbors. They did not measure health outcomes; therefore it is not possible to know whether the resulting decrease in

<sup>4</sup> The authors mention that their estimated valuation does not vary with the intensity of usage of the source at baseline (see footnote 12 in [Kremer et al., 2011a](#)), which could be indicative that the benefit is not decreasing with the quantity of clean water.

open defecation had impacts on health despite the absence of a complementary water provision intervention.

Further research could usefully document the nature of any interactions between water provision, sanitation access, and improved hygiene both for beneficiaries themselves and in terms of local spillovers. A large-scale intervention and study will likely be necessary to precisely estimate impacts on the full range of relevant health outcomes, including infant mortality.

Another feature of the environment that is becoming an increasingly important factor in life expectancy even in disease-ridden countries is road safety. Road deaths are the leading cause of death for people ages 15 to 29 worldwide. 90% of the world's fatalities on the roads occur in low- and middle-income countries, even though these countries have approximately half of the world's vehicles ([World Health Organization, 2015](#)). [Habyarimana and Jack \(2011, 2015\)](#) estimate the impact of a pilot and subsequent large-scale interventions to encourage passengers of 14-seater minibuses (the ubiquitous form of public transport in Kenya) to "heckle and chide" the bus driver when his behavior compromises their safety, in particular, when overspeeding or overtaking without visibility. The encouragement came in the form of evocative messages on stickers placed inside the minibuses. They find a large impact on insurance claims for minibuses randomized to receive the stickers, of between one-quarter and one-third depending on the stickers used, as well a decrease on the average maximum speeds and average moving speeds of the vehicles. In contrast, radio programs with the same encouragement seem to have no impact the weeks they are aired. Together, these results suggest that overcoming collective action problems is particularly important for reducing dangerous road behavior, but nudges may need to be very salient to be successful. They are an extremely inexpensive way to save lives however: [Habyarimana and Jack \(2015\)](#) estimate the cost-effectiveness of the most impactful stickers to be between \$10 and \$45 per disability adjusted life-year saved.

## 5. DEMAND FOR HEALTH PRODUCTS AND HEALTHCARE

As discussed in [Section 3](#), health is an input: it matters for how productive one can be. It is also a direct component of well-being (a consumption good, in the terminology used by economists). Both of these are reasons for individuals to invest in their health and that of their children. Health investments include preventive behavior, from getting vaccinated to wearing a seatbelt to avoiding risky sexual contacts, as well as prompt treatment of illness episodes, and diagnosis and management of chronic conditions. In a standard model with no market failures, the demand for any such health input or behavior is a function of its benefits, its costs (both monetary and nonmonetary), as well as the horizon over which both benefits and costs are accrued.

But many assumptions made in the standard model may not hold in low-income settings. Households in developing countries are often liquidity constrained, and they often lack information, or the education to process information, on the potential returns to various health investments. Even when information can be processed, judging whether it is correct may be difficult. As first noted by Arrow (1963), households' uncertainty about the true source of their health ailment makes learning about the quality of health care services, or of specific treatment or prevention tool, very challenging. This is particularly true for primary care where most illnesses are self-limiting, i.e., will get better over time (think of the common cold), and therefore signals are particularly noisy.

A second important issue for decision-making is widespread skepticism about outside advice, especially in societies where trust in the health care sector has been eroded for historical reasons, as in India where it is often argued that the forced sterilization effort carried out by Indira Gandhi's government during the "state of emergency" in the 1970s has created lingering distrust regarding government family health initiatives.

More generally, psychological factors are an important part of the health decision-making process for households in developing countries as well as in wealthy societies. Thinking about one's own mortality or the mortality of loved ones is unpleasant. People may thus tend to push such thoughts out of their mind unless there is no way to avoid them, such as when there is an acute illness episode or when an outside event makes health concerns highly salient. This has important implications for the design of health programs. A given intervention could have different impacts depending on whether it is implemented when health is at the top of people's mind or not.

There has been a large increase in the number of randomized experiments aimed at understanding the role of the more standard "economic" constraints and their implications for public health policy. Less research has been carried out on the trust and psychological factors directly, although considering those factors is useful for interpreting the results of existing experiments that act on the economic environment, as we discuss below. For example, the timing of an intervention relieving household liquidity constraints (e.g., whether or not it takes place while the household is suffering from an acute illness) could matter greatly for the take-up and impact of such an intervention.

In this section we review field experiments on the demand for healthcare and health products, sorting them not by the outcome they are examining but by the behavioral factor that they focus on, e.g. price factors, nonmonetary costs, cash on hand, etc. Because field experiments are often designed to study a number of such factors at once, often with a multiplicity of randomized treatment arms, our organizational scheme implies that we sometimes discuss the results of a given experimental study in multiple subsections. For other excellent recent reviews that are organized by health outcome (or sector) rather than behavioral factors, we refer the reader to Ahuja et al. (2010) for a review of

randomized evaluations on safe water access and [Jayachandran \(2015\)](#) for a recent review on gender inequality.

## 5.1 Pricing experiments

A number of field experiments have examined the role of prices in the adoption of health products and services. They typically do so by randomizing the price at which a household can access a product, and comparing take-up across price points, thereby tracing out the demand curve and estimating the price elasticity at different price points. The price elasticity is an important parameter because private health investments often have social externalities. Identifying that private demand is low may therefore justify government subsidies. For subsidies to not be wasteful, however, they have to strike a delicate balance: they have to maximize the likelihood that a needy person can access the health products or services that could benefit him or her, while also minimizing the likelihood that the subsidy accrues to those for whom the returns to the subsidy are inframarginal, either because they would have invested in the product privately anyway, or because they are unlikely to make effective use of products they receive at a highly subsidized price. This is a serious concern theoretically since households that are unwilling to pay a high monetary price for a product may also be unwilling to pay the nonmonetary costs associated with daily use of the product, or may not actually need the product at all. Indiscriminate subsidies would then undermine the screening or allocative benefits of prices. What's more, subsidies could also reduce the potential for psychological effects associated with paying for a product, such as a “sunk cost” effect in which people, having paid for a product, feel compelled to use it (an issue we return to below).

Below we review about half a dozen field experiments conducted over the last 15 years that have shed light on these issues. Before we go into their details, we discuss the relative merits of the various methods used to estimate willingness to pay at different price points.

### 5.1.1 Methods to estimate the demand curve

The most commonly used willingness to pay (WTP) elicitation method outside of field experiments is stated WTP: people are simply asked how much they would be willing to pay for the product. The main problem with this measure is that it is not incentivized; therefore, respondents may not think hard enough before providing their answer. Different individuals may also interpret the question differently if not asked precisely enough: some may report what they would pay if they had access to credit; some may report a low WTP if they think their answer may affect future subsidy policies; and others may exaggerate their willingness to pay to please the survey enumerator, etc. For this reason, researchers have moved towards field experiments in order to observe the “true” demand at each price point. For this, two main methods have been used, TIOLI and BDM, which we now discuss in turn and compare to each other.

Take-it-or-leave-it, or TIOLI, experiments randomize the price that an individual faces, observing whether that individual actually purchases the product at that price or not. This is a straightforward revealed preference mechanism.

The BDM mechanism, named after theorists Becker, DeGroot, and Marschak ([Becker et al., 1964](#)), is an incentive-compatible elicitation mechanism with real stakes that can be used to elicit individual willingness to pay as follows. People are asked to state the maximum they would be willing to pay for a product, i.e., make a bid, and to put forward their bid amount. Then a price is randomly drawn from a known distribution, and those who had bid at or above the randomly drawn price have to use the money they had put forward to purchase the product at that price (they keep the balance if they were willing to pay more than the price); while those who bid below the price cannot purchase the product. This mechanism is incentive-compatible, that is, it is a dominant strategy for expected utility maximizers, since those who bid less than their true value risk failing to buy the product when the price drawn is low enough that they would in fact prefer to do so. Conversely, bidding above one's true value entails the risk of buying when the price is higher than one would actually be willing to pay.

[Berry et al. \(2012\)](#) discuss the merits of each method in detail and compares them in field trials in Ghana. TIOLI is straightforward to implement through door-to-door experiments, voucher distribution or retail-level subsidies. Importantly, TIOLI can be done in a way that allows people time to think through and save for the product, for example by having a fixed TIOLI price in place for a certain time period, or distributing vouchers redeemable for a given number of months, as in [Dupas \(2009\)](#). In contrast, BDM can only elicit immediate willingness to pay, unless it is done over credit contracts. But BDM has the advantage of telling us, for each individual in the sample, what their exact willingness to pay is, whereas TIOLI only informs us of the share of the sample willing to pay at least a certain price. Thus TIOLI studies generally require much larger sample sizes. They also cannot easily be used to test for heterogeneity in outcomes based on individual willingness-to-pay without additional experimental features, such as a second, surprise randomization, as in [Karlan and Zinman \(2009\)](#), a method subsequently applied in the health sector by [Ashraf et al. \(2010\)](#) and [Cohen and Dupas \(2010\)](#), two studies we discuss further below. In contrast, BDM generates randomized variation in access to the product *within each observed willingness to pay group*: conditional on being willing to pay a given price, whether the study participant gets to acquire the product is random, and this allows researchers to estimate whether the returns to owning the product are heterogeneous by underlying willingness to pay ([Berry et al., 2012](#); [Chassang et al., 2012](#)).

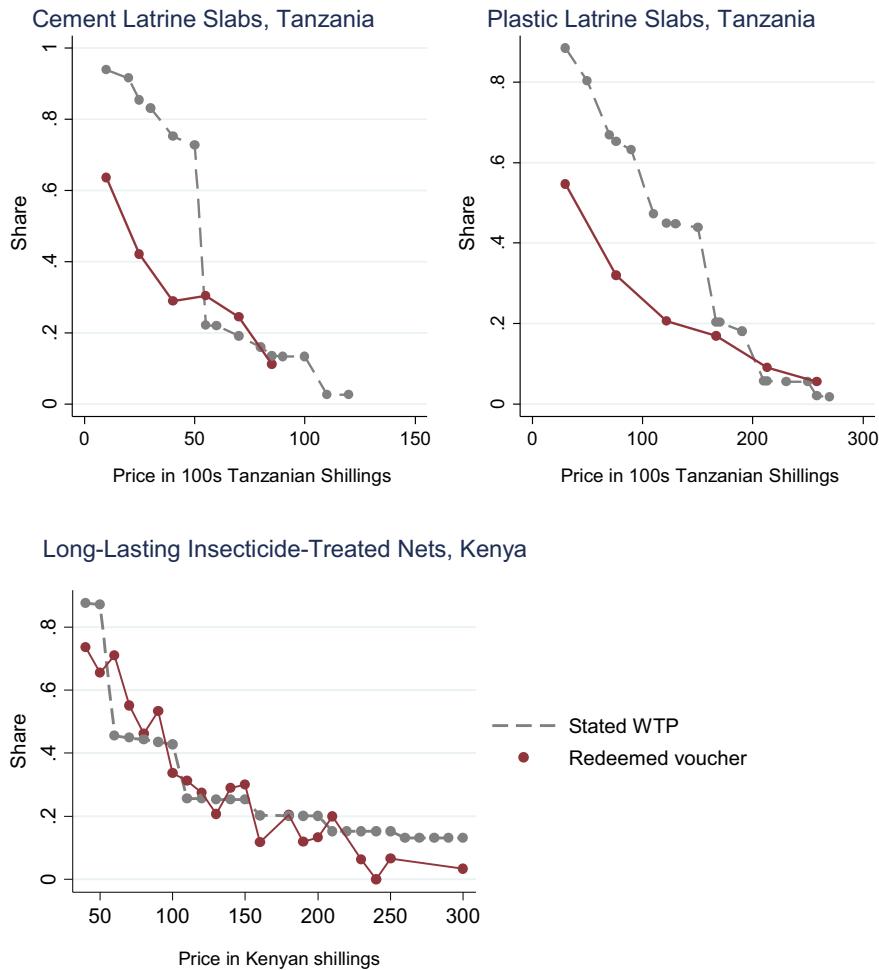
While BDM has the potential to generate richer data, the quality of this data is unclear, especially in resource-constrained settings where the population whose willingness to pay is elicited has low numeracy skills. [Berry et al. \(2012\)](#) assess the validity of the BDM mechanism in Ghana, comparing the demand curve for water filters obtained

through BDM with that observed through TIOLI (disabling the time feature of TIOLI, i.e., forcing people to decide on their TIOLI offer immediately). They find that even after shutting off the time dimension, BDM systematically underpredicts willingness to pay relative to TIOLI. The magnitude of this underprediction is not negligible, and appears to increase with price. Namely, the demand under BDM is 20% lower than under TIOLI at the lowest price considered (USD 1.40, a tenth of the retail price), 34% lower when the price is USD 2.80, and 45% lower when the price is USD 4.20. [Berry et al. \(2012\)](#) remain agnostic as to the reason why BDM appears somewhat inaccurate (if we take TIOLI as reflective of the “true” demand), though through additional experimental treatments, they can rule out that the difference between the two mechanisms is driven by either strategic bidding under BDM (i.e., people stating a low willingness to pay in the hope of influencing future prices, in particular the possibility of a greater NGO subsidy) or anchoring under TIOLI (if the TIOLI price influenced people’s willingness to pay). More work is needed to understand when and how the BDM approach can provide more accurate estimates of willingness to pay. In the meantime, most pricing experiments have been conducted using TIOLI.

As more pricing experiments take place, it should also become possible to better understand the circumstances under which stated willingness to pay, which is obviously far cheaper to elicit than either TIOLI or BDM experiments, gives a “sufficiently good” approximation of the demand curve. In [Fig. 1](#) we use the dataset underlying the experiment in [Dupas \(2009\)](#), as well as the dataset in [Peletz et al. \(2016\)](#), to compare stated willingness to pay to actual take-up in the TIOLI experiments. [Dupas \(2009\)](#) looks at the demand for bed nets, which are fairly well known products at the time of the study. [Peletz et al. \(2016\)](#) estimate demand for less common latrine slabs—concrete or plastic slabs with footholds and a hole for urine and feces, which can be used to cover latrine pits and are considered as health-enhancing as they reduce the risk of contamination of the home compound with fecal matter.

In both studies households were first asked at baseline how much they would be willing to pay for the product. In [Fig. 1](#), for each of the three products studied, the share reporting a WTP above each price point is plotted in grey dashed lines. In the subsequent TIOLI experiments, households in both studies had 3 months to redeem a voucher for the product at a specific price that was randomly varied across households. Only six price points were offered in the latrine slabs studies, while a wider range of prices was covered in the bed net study. The observed redemption rates at each price point are shown in solid maroon lines.

There are three interesting findings to note in [Fig. 1](#). First, there is a lot of “rounding” in stated willingness to pay: people tend to report amounts in multiples of 50 in Kenya and 5000 in Tanzania, which creates the illusion of very large drops in demand at certain price thresholds, when in reality demand is likely to be much smoother. Second, the accuracy of elicited WTP can be high, as in the bed nets case in Kenya, but in some other



**Figure 1** Observed take-up in TIOLI experiments versus stated WTP at baseline. (Courtesy of Peletz, R., Cock-Esteb, A., Ysenburg, D., Haji, S., Khush, R., Dupas, P., 2016. *The Supply and Demand of Improved Sanitation: Results From Randomized Willingness-to-Pay Experiments in Rural Tanzania* (In preparation) for slabs and Dupas, P., 2009. What matters (and what does not) in household's decision to invest in malaria prevention? *Am. Econ. Rev.* 99 (2), 224–230 for nets.)

contexts it seems quite inaccurate: in Tanzania where familiarity with latrine slabs was low, elicited willingness to pay vastly overstates demand. Third, there is little evidence of strategic underreporting of willingness to pay. This is particularly striking in the case of the slabs study, where the survey eliciting willingness to pay was carried out by NGO workers introducing themselves as interested in understanding how to improve health in the community, and yet stated willingness to pay was higher than observed take-up, likely due to wishful thinking regarding own take-up.

[Kremer et al. \(2011a\)](#) also estimate large discrepancies between stated preference valuations of access to a protected spring (elicited two different ways) and the revealed preference travel cost approach described in [Section 4](#), on the order of 2- to 5-fold differences. As discussed above, however, the travel cost approach relies on strong assumptions and therefore it is not clear what share of the discrepancy comes from stated preferences overestimating willingness to pay and what share of the discrepancy comes from the travel cost approach underestimating willingness to pay.

### 5.1.2 Results of pricing experiments

We now turn to reviewing over half a dozen randomized pricing experiments conducted to date. The great majority of those concern the pricing of *preventive* health products.

One of the earliest randomized TIOLI experiments for a health product in a poor country was, however, for a treatment product—a deworming drug—and took place in 2001 in Western Kenya. Among 50 primary schools enrolled in a free deworming program in 2000, [Kremer and Miguel \(2007\)](#) randomly selected 25 that moved to a *cost-sharing* program: parents now had to contribute a fee in order for their children to receive the deworming pill(s) on deworming day. Parents had to pay the fee at the school in advance of the deworming day, and were informed of this fee one or 2 months prior to treatment day. The researchers found that the share of children receiving deworming medication on the day the NGO visited the school for mass deworming was only 18% in the cost-sharing schools, compared to 75% in the schools who kept the free program, despite the fact that deworming remained heavily subsidized, i.e., the fee charged per child was just 20% of the actual program cost on average. Interestingly, parents of sicker pupils were no more likely to pay for deworming drugs, suggestive no screening effect of the cost-sharing program. While these results suggest that demand is highly sensitive to price, understanding why it is the case in this specific context is somewhat difficult. It could be that parents had gotten “used to” the free program and resented the introduction of the cost-sharing fee, and therefore their demand was lower than what it would have been had a free program never been implemented in the first place. It could also be, as the authors hypothesize is the case, that the perceived private value of deworming is lower than the fee charged, perhaps due to the treatment externalities that they document. Subsequent pricing experiments have adopted more nuanced designs in order to disentangle these mechanisms from each other.

[Cohen and Dupas \(2010\)](#) use a two-level randomized TIOLI design to estimate: (1) the demand curve for a new health product in rural Kenya: long-lasting antimarial bed nets (LLINs); and (2) the distinct roles of the screening and psychological sunk cost effects that price may have on their usage. LLINs cost \$7, and they prevent bites from malaria-carrying mosquitos while sleeping. The experiment, conducted in 2007, randomized the price at which prenatal clinics offered nets to pregnant women. Clinics charged either

0 (free distribution), 15, 30 or 60 US cents (note that the highest price point considered, 60 US cents, still represented a 92% subsidy).

This first level of randomization, at the clinic level, involved only 20 observations (20 clinics), something which has implications for inference, as discussed in [Cohen and Dupas \(2010\)](#). The second level of randomization was at the individual level. Namely, a random subset of women who had agreed to purchase the net for 30 or 60 cents was subsequently given a surprise rebate right after they had given their payment to the clinic's cashier. [Cohen and Dupas \(2010\)](#) find that demand is very sensitive to price: the likelihood that pregnant women acquired a net fell from 99% to 39% when price increased from 0 to 60 US cents (with the demand at the intermediate price points of 15 and 30 US cents at 92% and 72%, respectively). This suggests that while there is no discontinuity at zero (it was not the shift from free provision to any positive price that makes demand drop, but rather larger price increases), demand is on the whole quite price sensitive, with very low demand rates at prices that are still heavily subsidized.

They do, however, find that the rate at which pregnant women used the net (measured through home observation visits 2 months after distribution) was relatively high (60%); and it was completely independent of the price they had paid for the net, whether initially or after the surprise rebate. This suggests that there is neither a screening nor a sunk cost effect of prices in their context. Thus coverage (the share of pregnant women sleeping under a bed net), and hence its potential for public health outcomes, increases very rapidly as the price goes down.

In another TIOLI experiment conducted with a sample of households with school-aged children, also in Kenya, [Dupas \(2014a\)](#) found that demand for LLINs becomes slightly less price sensitive if subsidies are provided in the form of vouchers that households have 3 months to redeem at local retail shops: the demand at \$0.60 becomes 73%. But overall price remains the primary driver of the demand, with the purchase rate dropping to just around 33% when the price reaches \$1.50 (still an 80% subsidy) and to 6% when the price reaches \$3.50 (corresponding to a 50% subsidy). Various marketing strategies (e.g., making the morbidity burden or treatment costs salient, targeting mothers, eliciting verbal commitments to invest in the product) failed to change the slope of the demand curve ([Dupas, 2009](#)). But here again, the price paid did not matter for subsequent usage. In fact, home observation visits showed that usage of bed nets acquired through a subsidized voucher was extremely high, rising from 60% at a 3-month follow-up to over 90% after 1 year, and that was the case across all price groups, including recipients of fully subsidized nets. A similar level of bed net usage (90%) irrespective of initial price paid was observed in rural Zambia ([Fink and Masiye, 2015](#)), suggesting that this result is at least somewhat general.

The finding that demand is price sensitive has been established by TIOLI experiments for products other than deworming drugs and bed nets. In 2010, [Meredith et al. \(2013\)](#) randomized the subsidy level households faced for rubber shoes to prevent worm

infections in Kenya, and in 2008 they randomized the price of soap and vitamins in Uganda, Guatemala, and India. In all contexts, they found that demand was very sensitive to price. In a TIOLI experiment in urban Zambia, [Ashraf et al. \(2010\)](#), also find that demand for a bottle of water purifying solution (diluted chlorine) is sensitive to price, dropping from around 80% at the price point of 9 US cents (a 62.5% subsidy) to only about 50% at the full market price of 25 US cents. That experiment also used a two-stage randomization design but in this case both randomizations took place at the household level. Specifically, not just the surprise rebate but also the initial offer price was randomized across households. Using this design, they test for both the screening and sunk cost fallacy effects of prices. As in [Cohen and Dupas \(2010\)](#), they found no evidence of a use-inducing sunk-cost effect, but found some evidence of a screening effect of prices. Specifically, those households who had selected themselves into paying a higher price were more likely to have used the purification solution within 6 weeks of acquiring it, while those who had received a higher subsidy were more likely to still have it on their shelf, possibly because they were keeping it for later or, as per the authors' interpretation, because they were less likely to ever plan to use it for a health purpose.

The studies discussed above suggest that price is often not a good mechanism to target subsidies for health prevention tools to those who most need them. If anything, higher prices seem to create too many errors of exclusion, and to prevent the positive spillovers on disease transmission that may justify subsidies in the first place.

The evidence regarding health *treatment* products is somewhat different, however. [Cohen et al. \(2015\)](#), in a TIOLI experiment conducted in 2009 in the same area of Kenya as the bed net studies mentioned above, find that price can be (to some extent) used as a targeting mechanism to allocate malaria treatment. Targeting of malaria treatment is very important because of the negative spillovers that the overuse of antimalarials can generate: it can delay or preclude proper treatment for the true cause of illness, waste scarce resources for malaria control, and may contribute to drug resistance among malaria parasites, making treatment of malaria harder in the long-run. The reason why, within essentially the same population, price can be effective at targeting treatment when it is not effective at targeting prevention products (like bed nets) is that demand for treatment appears much less price-sensitive (especially among the poor) than demand for prevention. What's more, conditional on experiencing malaria-type symptoms, adults are much less likely to be malaria positive than children, but as with most treatments, the price per antimalarial dose for adults (who need to take more pills) is higher than the price for children. Consequently, at a given price per pill, children (the key target for the subsidy in this case) are on a flatter portion of the demand curve.

In addition to furthering our understanding of how price can be used to target health products in the developing world, that study makes two other contributions: (1) it highlights the trade-off inherent to subsidies for medications in environments with weak health system governance (which prevents conditioning the subsidy on a formal

diagnostic test result), and (2) it points out that bundling subsidies for medications with subsidies for diagnostic tests has the potential to improve welfare impacts, a point we come back to in [Section 5.3](#) below when we discuss information experiments.

While the studies above have mostly focused on the effect of price on contemporaneous demand, some field experiments have been specifically designed to look at the dynamic effects of prices. The questions here are the following: Can a one-off subsidy be enough to trigger learning and generate sustained adoption? Or is there a risk that people are unwilling to pay for a product they once received for free, as a cursory look at the [Kremer and Miguel \(2007\)](#) deworming cost-sharing results could suggest? This could happen if people, when they see a product being introduced for free, start to feel “entitled” to receive the product for free (i.e., they would “anchor” around the subsidized price).

To gauge the relative importance of these effects, [Dupas \(2014a\)](#) examines the long-run effects of the one-time bed net subsidy vouchers households received in the study mentioned above. Specifically, the research team came back a year after the first pricing experiment had been done, and implemented an additional stage in the study in which all households received a second subsidy voucher, but this time they all faced the same price of \$2.30 (a 70% subsidy). By observing how the take-up rate of the second, uniformly-priced bed net varies as a function of the price a household faced in the first year, [Dupas \(2014a\)](#) can test whether being exposed to a large or full subsidy in the first year (which, as discussed above, considerably increased adoption at that time) reduces or enhances willingness to pay for the bed net a year later.

Dupas finds that a larger initial subsidy *enhances* willingness to pay for a bed net a year later, suggesting the presence of a positive learning effect which dominates any potential anchoring effect. Interestingly, the learning effect trickles down to others in the community: households facing a positive price in the first year are more likely to purchase a bed net when the density of households around them who received a free or highly subsidized bed net is greater. Though once bed net ownership is widespread, the transmission risk starts to decrease and the returns to private investments decrease, and accordingly those with more subsidized neighbors in year 1 were less likely to invest in year 2. [Dupas \(2014a\)](#) also tests for cross-product effects of price subsidies, namely, whether getting a subsidy for the bed net led households to expect subsidies for a different health product, namely, a water purification product. She finds no such effect: willingness to pay for the water product a few months after being exposed to a subsidy for the bed net was not lower among households who had received a full or very high subsidy.

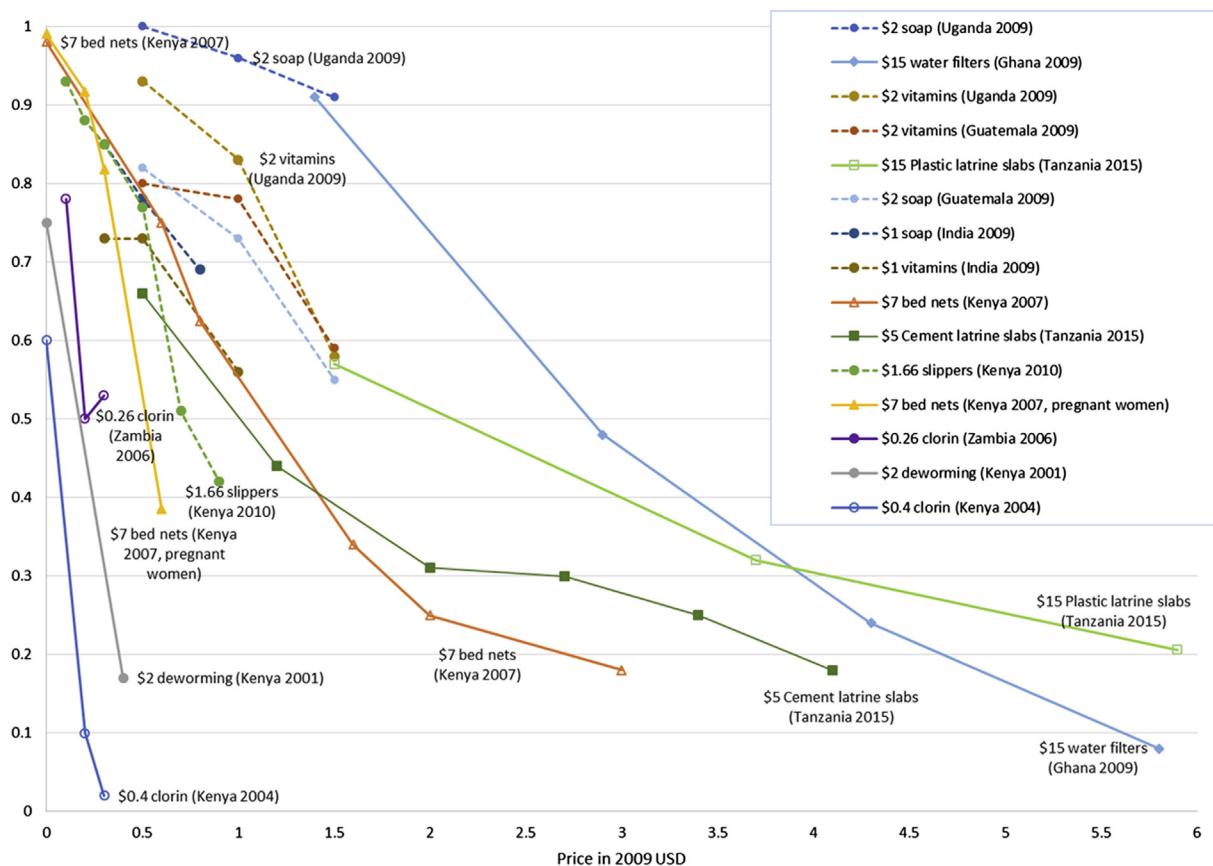
[Karlan et al. \(2014\)](#) adopted a similar research design to study the relative importance of learning versus anchoring effects in free distribution programs for a wider set of products, as well as the importance of who implements the free distribution. They conducted their experiment in northern Uganda. In a first round of door-to-door visits, they offer households one of three products either for free or for sale at the prevailing market price. The three products were chosen to differ in their scope for learning, and included a pain

reliever widely known at baseline (thus with little to no scope for learning); a deworming drug that was moderately well-known and which has some side effects, so that the learning effect, as in [Kremer and Miguel \(2007\)](#), might be expected to be negative; and a new, largely unknown treatment for childhood diarrhea for which the authors expected positive learning. The second door-to-door visits took place 2 to 3 months later and were conducted by a different set of sales agents. Households were offered the same product in round 2 as they had been offered in round 1, except for some randomly selected households who were offered a fourth product (a water purification solution) to test for the presence of any cross-product effects.

As in [Dupas \(2014a\)](#), [Karlan et al. \(2014\)](#) find no evidence of cross-product effects, but they argue that the patterns of demand they observe in round 2 are consistent with anchoring playing an important role: when there is no scope for learning (i.e., the pain killer case), they find that demand is lower after free distribution than after a sale. This is the case irrespective of whether the free distribution was done by an NGO or by a for-profit firm advertising the free distribution using a standard marketing tool (a “free trial” to help people learn a product is worth their money). The results are somewhat weak, however, with no significant differences in the treatment effects across products. One potential concern with the design of this study was that the authors did not collect information on the demand for the three products outside their experimental door-to-door visit, even though those products were available for a similar, if not lower, price at local shops.

This highlights a challenge facing many recent health demand studies: measuring impacts often requires measuring demand both within and outside the experiment. A good example of that is [Cohen et al. \(2015\)](#) discussed above, which measured access to Artemisinin Combination Therapy (ACT, the latest and most effective antimalarial) through not only the drug shops involved in the study but also local health facilities, something critical in their context given the scope for crowding out. This can be difficult if recall bias is a concern, especially when the time lag between baseline and endline is long. An alternative is the design used in [Dupas \(2014a\)](#), who offered a product (a long-lasting insecticide-treated bed net) unavailable on the market at the time of the experiment. Having perfect control over the supply means that observing the demand in the experiment provides [Dupas \(2014a\)](#) with a complete picture of the demand in both years. The drawback of that design is that when the product is not available on the market, the option value of experimenting with the product in round 1 may be lower, as households were not aware they would have a second chance to obtain the product from the experimenters, thus the take-up in round 1 may have been an underestimate of the take-up that would have prevailed in a real-world market environment.

[Fig. 2](#) graphically presents the TIOLI-estimated willingness to pay at various price points for 15 country–product combinations. For all products shown, the price points shown are at or below the market price. We indicate the country and year during which



**Figure 2** Purchase rate of preventive health products, by TIOI price. (Courtesy of Cohen, J., Dupas, P., 2010. Free Distribution or cost-sharing? Evidence from a randomized malaria experiment. *Q. J. Econ.* 125, 1–45; Dupas, P., September 12, 2014b. Getting essential health products to their end users: subsidize, but how much? *Science* 345 (6202), 1279–1281 for bed nets; Kremer, M., Miguel, E., 2007. The illusion of sustainability. *Q. J. Econ.* 122 (3), 1007–1065 for deworming; Ashraf, N., Berry, J., Shapiro, J., 2010. Can higher prices stimulate product use? Evidence from a field experiment in Zambia. *Am. Econ. Rev.* 100 (5); Kremer, M., Miguel, E., Mullainathan, S., Null, C., Zwane, A., 2011b. Social Engineering: Evidence From a Suite of Take-up Experiments in Kenya, Mimeo, Emory University for chlorine; Meredith et al., 2013. Keeping the doctor away: experimental evidence on investment in preventive health products. *J. Dev. Econ.* 105, 196–210 for soap, vitamins and slippers; Berry, J., Fischer, G., Guiteras, R.P., 2015. Eliciting and utilizing willingness-to-pay: evidence from field trials in Northern Ghana (CEPR discussion paper no. DP10703) for water filters; Peletz, R., Cock-Esteb, A., Ysenburg, D., Haji, S., Khush, R., Dupas, P., 2016. The Supply and Demand of Improved Sanitation: Results From Randomized Willingness-to-Pay Experiments in Rural Tanzania (In preparation) for latrine slabs.)

the TIOLI experiment was done for each product in the legend, as well as the market price. There are six broad types of preventative health products: soap, vitamins, bed nets, water filters, latrine slabs, and chlorine-based products to purify water.

Putting all of these results together in one figure yields a number of interesting patterns. First, and most importantly given that this is often misunderstood (see, for example the inaccurate treatment of this literature in the “Health Chapter” of the 2015 World Bank Development Report, [World Bank, 2015](#)), demand at small but nonzero prices is often substantial: at the USD 0.50 price point, demand is over 65% for 9 out of 13 country-products for which this price point was included. At the USD 1.5 price point, it is over 50% for 6 out of 13 country-products (the price of chlorine is below USD 1.50 so we exclude the two chlorine studies from the denominator). This is important to note because many earlier review papers have been interpreting the evidence from pricing experiments as suggesting a sharp drop as soon as the price is not zero, but this is definitely not the case. In fact, there are only two products for which there is a sharp drop in demand as soon as the price is not zero: this was documented for deworming in Kenya in 2001 ([Kremer and Miguel, 2007](#)) and for chlorine in Kenya in 2004 ([Kremer et al., 2011b](#)). In the deworming case, the authors show that the return to private investment in deworming is possibly negative given the large externalities, and possibly in part due to the side effects associated with treatment. More speculatively, parents needed to make the payment for their children to be dewormed at the school in advance of the treatment day, so one potential contributor to the low take-up could be a lack of trust that the payment would indeed be followed by treatment.

In the chlorine case, the striking fact is that demand is only 60% even at zero price. This suggests that chlorine is a product for which a substantial portion of the population has little to no demand, possibly owing to the residual taste it sometimes leaves in water. The fact that small prices lead to a large drop in demand could come from the fact that many people who do not value the product because of the taste still take it if it is free in order to use it for other purposes (e.g., cleaning), hence the drop from zero to a positive price may not directly reflect a drop in health usage. Or it could be that people take the first free sample and once they have learned they do not value the product, they stop taking it even if it is free (indeed in [Kremer et al. 2011b](#), demand for chlorine at nonzero prices was observed 2 months after everyone had received a free sample bottle for free).

While there may not be anything “special” about the price point of zero in many cases, it is evident from [Fig. 2](#) that investment in preventive health is highly sensitive to price even when the price is below the prevailing market level. Moreover, the evidence from these studies concerning the characteristics of those who select into paying higher prices suggests that prices are usually not a very effective allocation mechanism in the sense that they fail to target those who appear to need the products the most (see [Dupas, 2014b](#); for a review).

[Kremer and Glennerster \(2011\)](#) present a framework highlighting that the price sensitivity documented above may be due to liquidity constraints, lack of information, nonmonetary costs, or behavioral biases such as present bias and limited attention. A number of experiments have generated randomly varying access to liquidity, convenience or information, sometimes interacted with random variation in prices, to estimate the relative role of these potential factors. The evidence to date suggest that information about a product is necessary but not sufficient, and in particular information does not appear to substitute for higher subsidies, while reducing nonmonetary costs and increasing liquidity often matter a lot for take-up. We review this evidence below, before discussing their implications for the scope of behavioral factors at play in [Section 5.8](#).

## 5.2 Liquidity experiments: credit and cash transfer experiments

In the pricing experiments discussed above, households commonly report lacking access to credit as a reason for not being able to take up the subsidized products offered. To test whether such reported liquidity constraints are indeed a serious barrier, one would need to allow a random subset of households to purchase health products on credit. While researchers who have exploited the random introduction of microcredit have typically not found impacts on health expenditures (see [Banerjee et al., 2015d](#) and references therein), this may due to the coarseness of their data on health investments, and/or to the fact that most microfinance institutions focus on business loans rather than consumption loans, and that “flypaper effects” (the fact that money sticks where it first “hits,” like a fly on a flypaper) are common ([Fafchamps et al., 2014](#)).

The first studies (to our knowledge) to directly study demand for health products at full price when credit constraints are relaxed are [Devoto et al. \(2012\)](#) and [Tarozzi et al. \(2014\)](#). [Devoto et al. \(2012\)](#) identified low-income households not connected to the water grid in the city of Tangiers in northern Morocco, and randomized which households were told about a credit program to purchase a water connection. They see an impressive take-up rate of 69%, despite the fact that the cost of the connection (which varies with distance to the water mains) averages around \$1000, an amount that they would have to repay over 5 years. [Tarozzi et al. \(2014\)](#) randomized access to ITNs on credit across villages in the state of Orissa, India. They find that 52% of households offered full-price ITNs on credit purchased at least one ITN (and all of them fully repaid the loan). In contrast, in a follow-up cash sales study, they find that only around 11% of households purchase at least one ITN in the absence of any credit. [Fink and Masiye \(2012, 2015\)](#) also examine the demand for full-price bed nets when people are offered a zero-interest loan in the context of rural Zambia. They find that households offered bed nets on credit acquire 0.8 nets on average, a demand comparable in magnitude to that observed in [Tarozzi et al. \(2014\)](#).

Another set of studies looks at the impacts of cash transfers on health choices. In their pricing experiment described in Section 5.1, [Meredith et al. \(2013\)](#) gave households a randomly determined amount of cash (cash drops in the form of payouts for incentivized risk preferences elicitations) at the same time they distributed discount vouchers for rubber-sole shoes aimed at protecting children from worm infection. The market price for the shoes was about 85 Ksh (\$1.13) at the time, and discount vouchers varied from a low (20 Ksh) to a high (80 Ksh) discount. The cash drop varied from 0 to 200 Ksh, with a mean of 35. The researchers use the variation in cash drop amount to estimate the effect of liquidity on purchase.

They find a large and significant effect of the cash drop on demand: on average, every additional 100 Ksh in randomized cash payout increases the probability of voucher redemption by 22 percentage points. Since the cash drop was very small relative to lifetime income, at about 4% of weekly income on average, its effect on demand reflects a cash-on-hand effect rather than an income effect. Importantly, since households had to travel to a local store to redeem a voucher in order to obtain the product, the cash drop effect is arguably unlikely to be driven by an experimenter demand effect (whereby study households would feel compelled to use the money obtained from the experimenter to purchase the product sold by the experimenter). That said, some form of experimenter demand effect cannot be ruled out, as it may have been in the implicit priming embedded in the subsidy. The subsidy may have made health particularly salient in participants' mind, and that is why they chose to spend the extra cash on a health product, rather than something else.

Indeed, in contexts where individuals are not primed to think about health, the impact of cash drops on health investments has been mixed. [Banerjee et al. \(2015c\)](#) find that a “graduation program” that combines the transfer of a productive asset with consumption support, training, savings encouragement and basic health services to very poor households can lead to sustained income improvements for these households, but they do not find particularly meaningful impact on physical health despite the income change. It is important to note that their data comes from six countries with quite different underlying health burdens, however, and the researchers do not present data on health investments tailored to each context (e.g., bed nets for Ghana or water filters for India).

[Haushofer and Shapiro \(2013\)](#) find that provision of fairly large unconditional cash transfers to households in Kenya does not affect the health of under-5 children as measured by their height, weight, and upper-arm circumference, but here again they do not report any information on investments in preventative health behavior such as bed nets for adults (since at the time of the experiment under-5 children were supposedly covered by a government distribution scheme), latrine upgrades, etc. In contrast, in an experiment that primed people to save for health, [Dupas and Robinson \(2013\)](#) find

that investment in preventive health increased significantly for those who gained access to a simple saving technology.

One study focuses more specifically on the reproductive health impacts of providing cash transfers to adolescent girls in Malawi. [Baird et al. \(2012\)](#) randomly assigned 175 enumeration areas (EAs) to three groups: girls in 46 EAs received conditional cash transfers (CCTs) if they achieved 80% school attendance; those in 27 EAs received unconditional cash transfers (UCTs); 88 EAs served as a comparison group and did not receive transfers. The cash transfers significantly lowered the prevalence rates of HIV and the herpes simplex 2 virus (HSV2). For example, 1.2% of girls enrolled in school at baseline who received transfers (CCT or UCT) tested positive for HIV at 18 months relative to 3% of girls in the comparison group. Self-reported sexual behavior was also lower among girls who received transfers; 3% of girls who received transfers reported having sex at least once per week, compared to 7% in the comparison group. These results suggest that financially empowering school-aged girls can have substantial effects on their sexual and reproductive health choices. Interestingly, the authors find that the amount of the money transferred did not itself matter, nor the share of the transfer directly transferred to girls vs. their parents.

### 5.3 Information experiments

Even when liquidity constraints are alleviated, adoption of high-return health products or behaviors is often not 100%. A potential explanation for this could be that individuals' lack of information on the health costs or benefits of different products or behaviors. In this section, we review information experiments showing that (exhaustive) information is necessary but often not sufficient to generate take-up.

#### 5.3.1 Impact of information on willingness to pay

In their antiworm rubber-sole shoes pricing experiment in Kenya, [Meredith et al. \(2013\)](#) find that health workshops did not affect total demand nor the price gradient in demand. In their complementary evidence from India, Guatemala and Uganda, they find an effect of the health information in only one of six country–product combinations they experimented with—namely, a health script delivered at the time households could obtain subsidized soap in India flattened the effect of price on demand. The authors discuss that this one result may be driven by an experimenter demand effect, however, since in this specific case the purchase decision was contemporaneous with the cash drop.

In China, [Ma et al. \(2014\)](#) find that information and training on the importance of wearing eye glasses for children needing correction had no impact on take-up in the absence of subsidies, but a significant impact on usage of free eye glasses. Their context is the following: While eye examinations with close to 20,000 primary school students suggested that about 16% needed glasses, only 15% of those had glasses at the time of the baseline. The study distributed free eyeglasses to a random subset of those who

needed them (the randomization was done at the school level), and find that usage (observed through unannounced spot checks in the classroom) doubles among those diagnosed as needing glasses—but it remains stubbornly low at just around a third. A fairly intense training program (showing a short documentary-type film, handing out a set of cartoon-based pamphlets for students, and a lecture and handout for parents and teachers) had no impact on eyeglasses wearing in the pure control group (without free distribution) but was found to increase usage in the free distribution group by 17 percentage points or 30% compared to free distribution alone.

[Ashraf et al. \(2013\)](#) also interact the subsidy level and information provision, but the information provided concerns the relative merits of one product over another, rather than absolute information about the returns to preventing or treating a condition compared to doing nothing. Specifically, in a door-to-door marketing campaign in urban Zambia, sampled households were offered the option to buy one of two water purification products, a product well known in the area and available at retail stores (called “Clorin”) and a similar product from another brand, which people had never seen before and which was not available at any local stores. The price of the familiar product was fixed at 800 Kwacha, the standard retail price. The price of the unfamiliar product was randomized across households, and varied from 0 to 1200 Kwacha. In addition to the price randomization, the information given about the unfamiliar product was randomly varied: half of all households were provided no information, while the other half were told that the unfamiliar product is similar to and as effective as the familiar product.

[Ashraf et al. \(2013\)](#) find no overall impact of the information treatment on the demand for the new product, or on total demand. However, the demand curve for the unfamiliar product becomes steeper when information is provided, and it pivots exactly around the price at which the familiar product is available. Demand for the familiar product (as a function of the price of the unfamiliar product) pivots the opposite way, and total demand for water purification products does not increase significantly in the presence of information. This apparent complementarity between information and subsidies can be interpreted as follows: in the absence of any information, people tend to take the price of the unfamiliar product as a signal of its quality, so they are not completely turned off by high prices, while they are somewhat turned off by low prices. When information is provided, the signaling content of the price diminishes. As a result, people are less likely to be turned off by low prices, and more likely to be turned off by high prices (in particular, there is now no reason why they would pay more for the unfamiliar product than the price of the familiar product, since the information reveals that the two products are comparable). The effect of information provision is thus to encourage more people to switch from the familiar product to the unfamiliar product at low prices, and to deter more people to do the switch at high prices.

On the whole, the few existing studies examining the impact of information on willingness to pay have found limited impacts on total level of investments in health

products, but suggest that the impact of subsidies on health can be heightened when the subsidy is accompanied with information.

### **5.3.2 Impact of information on health behavior change**

A number of experiments have studied the impact of information on health behavior change. In Kenya, as part of the provision of deworming medication discussed above (in [Miguel and Kremer, 2004](#)), [Kremer and Miguel \(2007\)](#) randomly varied whether schoolchildren received information on how to avoid intestinal worm infections. The information was provided in the classroom by a mixture of trained teachers and NGO staff, and focused on preventative behaviors such as washing hands, wearing shoes, and avoiding infected fresh water. One year later, data on pupil cleanliness and shoe wearing (as observed by the research team) as well as self-reported data on exposure to fresh water showed no effect of the education campaign.

Also in Kenya, [Duflo et al. \(2015a\)](#) and [Dupas \(2011b\)](#) examined the impact of providing different types of HIV/AIDS information to primary school students. In a randomly selected subset of 328 schools, teachers were trained on how to implement the national HIV/AIDS curriculum, which focuses on abstinence as the only prevention method available for adolescents. [Duflo et al. \(2015a\)](#) find that the training greatly increased the likelihood that teachers teach about HIV in the classroom, and 2 years after the training students whose teachers had been trained had greater knowledge about the disease. The intervention did not reduce childbearing rates among girls, however, suggesting that it did not decrease the likelihood that girls engaged in unprotected sex. It also did not reduce the risk of STI as measured after 6–7 years.

Within the 328 schools, [Dupas \(2011b\)](#) randomly selected a separate 71 schools to receive an information session that discussed the role of cross-generational sex in the spread of HIV, and the relative risk of HIV infection by gender and partner's age. In many African countries, HIV prevalence increases with age among men. Therefore sex with older partners, which in many cases occurs in relationships with so-called "sugar daddies," substantially increases the risk of HIV infection for adolescent girls. This information was provided by a trained facilitator, introducing herself as working for a local NGO, to upper grade students in the selected 71 schools. This "relative risk" information intervention, which provided adolescents with information on how to reduce their exposure to HIV conditional on being sexually active rather than only exhorting them to abstain, led to a 28% decrease in teen pregnancy among school-going adolescent girls, and was driven by a reduction in cross-generational sex (with male partners five or more years older). Together, the results of these two experiments suggest that providing specific information is more effective than general exhortation at changing sexual behavior.

Rather than experimenting with the *content* of the information provided, two recent studies have experimented with the *delivery method* for HIV information. Indeed, recent advances in communication technology mean that information does not need to be

delivered in-person by either a teacher or an outside facilitator. In Colombia, [Chong et al. \(2013\)](#) looked at the impact of an online sexual health education course provided through schools. Researchers partnered with Profamilia, a large Colombian NGO, to randomly provide a course to one-third of 138 ninth-grade classrooms from 69 public schools in 21 cities. One-third of the classrooms were randomly assigned to the comparison group, which did not receive the program, while the remaining one-third of classrooms did not participate in the course but were located in the same schools as the classrooms that did receive it. The course increased overall sexual health knowledge by 0.38 standard deviations, and increased positive attitudes towards condom use. There was no impact on self-reported sexual behaviors, but there was a reduction of 5.2 percentage points (83%) in self-reported sexually transmitted infections among females who were already sexually active before the program, suggesting that some students adopted safer sex practices. The reliance of the study on self-reported sexual behavior is somewhat problematic, however, for reasons discussed in [Section 2](#) above.

In Uganda, [Jamison et al. \(2013\)](#) tested the impact of increasing access to information about sexual and reproductive health for the general population (not just schoolchildren) via a text messaging service about risky sexual behavior. Among 60 villages, marketing teams encouraged individuals in a random subset to use a new mobile phone-based information system through which users could send questions and receive responses on sexual and reproductive health. Usage among these villages was fairly high at 40%, but the service had no discernible impact on villagers' sexual or reproductive health knowledge. The intervention did, however, lead to an overall increase in the incidence of risky sexual behavior and self-reported promiscuity, particularly among men, while women reported increased abstinence. Qualitative information sheds some light on the potential causes of this mixed impact: men and women both reported that married women who learned about the risks associated with having an unfaithful partner insisted their husbands be faithful and get STI tested. According to these reports, some husbands did not comply, leading women to deny them sex and men to seek sex from alternative partners. Overall, individuals in treatment villages perceived their sexual behavior to be riskier, which could indicate an actual increase in risky behaviors, or could indicate that the information service increased accurate assessment of health risks. Unfortunately the researchers here again do not have objective measures of the risk level, such as biomarkers of sexually transmitted infections or pregnancy, to tease out these two potential explanations.

In Malawi, [Godlonton et al. \(2016\)](#) estimate the impact on sexual behavior of an information campaign about the relationship between circumcision and HIV status. They study the impact on men who are not circumcised at baseline, as well as those who are. For the former group, they find that the information increased correct knowledge about relative risk, reduced risky sexual activity, and increased condom use. Specifically, uncircumcised men in the treatment group were 25% less likely to have sex each month and

58% more likely to use a condom. For the latter group, which learned they were better protected, there was no evidence of riskier sexual behavior. While uncircumcised men reported an increased willingness to have their male descendants circumcised, overall take-up of adult male circumcision was low. Researchers also found that the circumcision information campaign, though it increased correct understanding about how male circumcision can partially protect males against HIV transmission, also increased the incorrect belief among participants that male circumcision protects females against infection as well (which it does not). These results suggest that information alone is not enough to increase the demand for male circumcision, and that one has to be careful in the way information is delivered to mitigate the risk that incorrect beliefs are propagated.

Also in Malawi, Chinkhumba et al. (2014) conducted a randomized experiment to study the impact of information and prices on the demand for medical male circumcision. 1634 men were given vouchers for a subsidized circumcision at a nearby clinic; the researchers randomly assigned different values to the vouchers, with subsidies ranging from 8% to 100% (i.e., free) of the full price. The study randomly selected half of the men to receive comprehensive information about the biological relationship between male circumcision and HIV risk. Results were collected through both self-reports and clinic records. Information increased the number of circumcisions by 66% (1.4 percentage points), but overall the rate of circumcision was extremely low: no one offered the full price was circumcised, and only 3.1% of those offered a free circumcision elected to take up the procedure.

Banerjee et al. (2015a) compare the impacts of two information interventions on the take-up of salt fortified with both iron and iodine (Double Fortified Salt, or DFS) in rural Bihar, India. The extremely high rates of anemia in the region suggest that widespread adoption of DFS could have important health impacts for adults as well as children. In the absence of any information intervention, however, just under 10% of households use DFS about 2 to 3 years after the introduction of the product at a subsidized price in their villages (and only 20% have ever tried it). This suggests a potential role for information interventions.

The authors compare a light touch information intervention—namely, a flyer hand-delivered at respondents' homes, informing them of where DFS can be bought locally—to a full-fledged “infotainment” intervention. The infotainment intervention consisted of a specifically designed movie which lasted 26 min and was shown during the intermissions of two screenings of a classic Bollywood movie, first at night in the center of the village and again the next day at a school or health center. The infotainment movie aimed to showcase the health benefits of adequate iron consumption and the availability of iron in DFS in an entertaining way. The movie was modeled on sitcoms and starred real actors from the local movie industry, and the main character in the movie is a short and scrawny man who dreams of having a tall and strong son.

During a prenatal care visit, his pregnant wife learns of the importance of taking iron supplementation for the health and future physical development of the unborn child, and convinces him to purchase DFS.

In light of the low baseline take-up in the absence of any intervention, the movie intervention had a large impact: at follow-up, conducted between 7 and 16 months after the movie was shown in the village, DFS use was 5.5 percentage points (57%) higher in treatment villages. Having “ever used” DFS was 11.5 percentage points (22%) higher. Based on observed viewership for a random subset of the screenings, the authors estimate that someone had seen the movie in roughly 20% of households on average, so the impact of the infotainment intervention in “per viewer” terms is very large. In contrast, the “light touch” encouragement (the home-delivered flyer) had no apparent effect on usage. This suggests that a heavy touch is needed for information on the importance of DFS to sink in.

Interestingly, in a parallel experiment the authors tested the importance of incentives for retailers in the diffusion of DFS. In the treatment arm for that experiment, retailers selling DFS were given financial incentives in the form of higher markups on the DFS. This intervention had an impact on take-up of the same magnitude as that of the infotainment, but not because retailers were effective at changing households’ perception of the importance of consuming DFS: retailers did not appear to try to convince households of the importance of DFS for health, or if they tried, they did not succeed. Instead, they pushed the DFS on households by claiming to have no other salt in stock. The fact that shopkeepers cannot be successful information messengers for new health products is perhaps not surprising, since their credibility as a neutral information agent is limited unless they can prove that they are not making more profit from selling DFS compared to other types of salt. That is to say, by instituting greater financial incentives to boost sales of a new health product will likely undermine the potential for shopkeepers to be effective knowledge agents about the product.

### **5.3.3 Impact of tailored information on behavior change**

All the studies above concern generic information. In contrast, [Prina and Royer \(2014\)](#) study the impact of providing tailored (individual-specific) information—namely, the impact of providing parents with body weight report cards for their own school-aged children. The report cards included information on a child’s height and weight as well as their weight classification (i.e., underweight, healthy weight, overweight, or obese). This intervention increased parental knowledge and shifted parental attitudes about children’s weight, but did not lead to meaningful changes in parental behaviors or children’s body mass index, even when the body weight information was accompanied by information on the health risk of obesity. The authors provide evidence that social norms matter: if the report card included information on the distribution of weights in the classroom, then the larger the fraction of overweight children in the child’s class, the less likely

a parent was to report that her overweight child weighed too much. As the authors note, this implies that as obesity rates increase, programs aimed at reducing obesity may become less and less successful, as local reference points for appropriate body weights may rise.

Also looking at the impact of tailored information, [Madajewicz et al. \(2007\)](#) test the impact of informing households in Bangladesh about the safe/unsafe status of the arsenic concentration of their local well water. They find that the information treatment increased water source switching: 60% of households informed that they were using unsafe wells changed wells, compared with only 8% of households in control areas changing wells within the same time period. In contrast with [Dupas \(2011b\)](#) in the context of HIV risk information, [Bennear et al. \(2013\)](#) show that adding to such bright line message (“this well is safe/unsafe”) some information about the dose response of arsenic contamination (namely, the fact that switching to a well with a lower level of arsenic concentration is always better than not switching, even if the switch-into well is not below the national “safe” standard) does not help—it does not increase switching, in fact if anything it decreases it, although the authors cannot reject the null hypothesis of no additional effect of the dose response information.

Diagnostic test results are another form of tailored information. Access to such tests is fairly limited in many developing countries where testing is expensive and hard to access. The experiment in [Cohen et al. \(2015\)](#) included randomized access rapid malaria diagnostic tests (RDTs) among households in Western Kenya, an area where overtreatment with malaria appears very common due to poor access to reliable diagnostic tests. They find that conditional on seeking care for presumed malaria illness, those who learned their malaria status through an RDT were 40 percentage points less likely to buy malaria medicine than those who did not know their status at the time of purchase, reflecting the fact that only 36% of adults seeking malaria treatment in response to a presumed malaria episodes do not actually have malaria. Still, they find that compliance with a negative test result is not perfect: just around half of adults testing negative for malaria still went on to purchase an antimalarial drug. This cautiousness in learning from one’s own test results exemplifies the complexities associated with learning in the very noisy health environment that many people find themselves in, where information on the reliability of the test itself is limited. In an environment with at least three unknowns—the true underlying cause of an illness episode, the relative efficacy of drugs given a cause, and the accuracy of diagnostic tests—establishing over time the reliability of information provided by local health “experts” is extremely challenging. Since many diseases are self-limiting, a true nonmalaria episode nevertheless treated with an antimalarial treatment may appear to benefit from the treatment (and may thus be misperceived as a false negative on the test) even though it would have resolved equally rapidly without treatment.

[Delavande and Koehler \(2012\)](#) and [Gong \(2015\)](#) exploit randomized access to HIV testing to study the impact of learning one’s HIV status on sexual behavior in Malawi,

Kenya and Tanzania, prior to the introduction of antiretroviral therapy. HIV risk and sexual behavior is another domain where there are a number of often unknown parameters—the status of partners, the transmission rate per unprotected act conditional on having sex with an infected partner, and one's own status. As a result and as previously argued by [Boozer and Philipson \(2000\)](#), the impact of HIV testing on subjective beliefs and risky behavior is theoretically ambiguous. Learning about one's status at a point in time provides a joint signal about the transmission risk and previous partners' statuses. Depending on priors regarding those, the HIV test result can thus lead individuals to update their beliefs about the transmission parameter in opposite directions. Those who know they had an infected partner in the past may revise downward their beliefs about the transmission risk. At the same time, a test result can change one's time horizon—in particular, assuming they trust the test enough, a negative test result would increase expected life expectancy for those who were pessimistic about their status before. In contrast, those who get a positive test results would suddenly have “nothing to lose” (especially in the absence of treatment).

[Delavande and Koehler \(2012\)](#) exploits the randomized access to HIV test results from the 2004 experiment by [Thornton \(2008\)](#), discussed in [Section 5.5](#), to look at some of these pieces of the puzzle in the context of Malawi. They find surprising results: those who learned they were HIV negative in 2004 are more pessimistic about their status in 2006 than those who did not learn their status, and also have less precise beliefs. Those who learned they were HIV positive did not change their beliefs about their own status, and if anything revised downwards their beliefs about the transmission rate (the authors suggest this could be because a number of them jointly learned their spouse was HIV negative). Ultimately they find no impact on risk taking among those who learn they are negative and a decrease in risky behavior among those who learn they are positive. They do not perform the analysis separately by baseline prior regarding one's own risk however, as they only have data on subjective expectations at follow-up. They also do not have objective measures of risky behavior.

[Gong \(2015\)](#) is able to overcome both of these shortcomings in his re-analysis of a randomized study of voluntary counseling and testing (VCT) for HIV, conducted in the mid-90s in Kenya and Tanzania ([The Voluntary HIV-1 Counseling and Testing Efficacy Study Group, 2000](#)). In this earlier study, baseline data on expectations were collected, as well as data on STI incidence over the 6 months that followed the randomized VCT intervention. [Gong \(2015\)](#) tests whether STI incidence at the 6-month follow-up varies as a function of treatment status among four subgroups: (1) those who are HIV- and anticipated they were; (2) those who are HIV- but thought they were not; (3) those who are HIV+ and anticipated they were; and (4) those who are HIV+ but thought they were not. The author finds significant impact of the VCT treatment (learning your true HIV status) among those who are “surprised” by the test result (people in groups 2 and 4). In particular, those surprised by a negative test result are less

likely to get an STI. While those surprised by a positive result are more likely to get an STI. The author argues that these differences in STI incidence are due to differences in sexual behavior—those surprised by a negative test result adopt safer behaviors while those surprised by a positive test result adopt riskier behaviors. In contrast, unverifiable self-reported sexual behavior data finds decrease in risky behavior for all groups.

While the set-up in [Gong \(2015\)](#) is very appealing, a weakness is that the sample is limited in size. What's more, it suffered from important attrition between testing and follow-up, and the subgroup analysis was an afterthought. In fact, the team of researchers who initially implemented the randomized study had estimated that “the study was not powered to detect significant differences between groups in the rate of incident sexually transmitted disease” ([The Voluntary HIV-1 Counseling and Testing Efficacy Study Group, 2000](#)). Thus one concern with [Gong \(2015\)](#)'s analysis could be one of ex-post data mining, since doing the analysis by subgroups based on priors was not prespecified.

### **5.3.4 Impact of targeted information**

A final question regarding the role of information in health behavior concerns whom the information should target. [Ashraf et al. \(2014\)](#) explore this question in the context of the demand for family planning in Zambia. Women in the study received vouchers that granted appointments with a family planning nurse at the local government clinic. Information explaining all methods of family planning, including “concealable” methods such as injectables, was provided along with the vouchers. Women were randomized into two treatment groups. In the “individual” arm of the study, women were given these vouchers alone. In the “couples” arm, women were given these vouchers in the presence of their husbands. In all other respects, the experimental protocol in the individual and couples arms was identical. The authors find that take-up of the voucher was significantly lower when information about family planning services was provided in the presence of husbands: women who received the voucher in the presence of their husbands were 9 percentage points (18%) less likely to use the voucher to obtain an appointment at a family planning clinic. This gap was larger (12 percentage points) among couples with divergent fertility preferences, in particular, where the husband reported wanting more children than the wife, strongly suggesting that individual health choices cannot always be easily separated from within-household bargaining issues or other forms of interaction among household members.

## **5.4 Schooling experiments**

Educational attainment is an important predictor of health, even conditional on income, in most countries, irrespective of their level of development ([Strauss and Thomas, 1995](#); [Cutler and Lleras-Muney, 2014](#)). While there are a number of potential channels through which years of schooling could have a causal impact on health—for example,

if educational attainment increases how much information people have or are able to process about the health impacts of various behaviors, or if it changes individual's discount factor or intrinsic valuation of health—disentangling those from the reverse causal channel, in which better health increases the incentive to invest in education, is methodologically difficult.

[Jensen and Lleras-Muney \(2012\)](#) follow up a randomized intervention that increased schooling among men in the Dominican Republic and estimate a reduction in the incidence of their heavy drinking and smoking in the short run. They argue that the effect came about mainly by changing subjects' resources and peers (since at the time of follow-up, those with more education were still in school and thus less likely to work).

While there is no randomized experiment conducted over a sufficiently long time-frame to provide causal estimates of the effect of education on health in the long run, evidence on medium run effects can be found in [Duflo et al. \(2016\)](#), who randomized access to secondary education in Ghana by offering scholarships to a random subset of students who had received admission into senior high school but had difficulty enrolling due to financial hardship. The scholarship, offered in January 2009, produced a large difference in enrollment and completion of secondary school: among boys (girls), 75% (65%) of the scholarship winners completed senior high school compared to 47% (36%) of the nonwinners. Since 2009, [Duflo et al. \(2016\)](#) have been keeping track of the 2064 youths in their study sample (a third of whom were randomized into the scholarship treatment) in order to study how this gap in educational attainment translated into gaps in later life outcomes, including health. Longer term effects on objective health levels are yet to be measured, but as of the 7-year follow-up they find significantly higher rates of preventive health behaviors, such as bed net use, condom use, handwashing with soap and use of mosquito repellent, among those who were offered the scholarship. A likely channel for this finding is the information channel: they find little evidence for price or income effects, as their sample is not yet well integrated in the labor market at the time of the follow-up, and also find no effect of education on individual discount factors. Scholarship winners scored an average of 0.16 standard deviation higher on a reading and math test administered in 2013, and were also more likely to engage with the media (e.g., read the newspaper), suggesting that schooling differences also led to cognitive skills gains, another potential channel.

## 5.5 Nonmonetary cost experiments

For some products or services, demand remains low, even at very low or even zero prices. For example, in the poor district of Udaipur in India, despite the fact that immunization services were offered for free in public health facilities, [Banerjee et al. \(2010\)](#) estimate that only 2% of children aged between 1 and 2 had received the recommended basic package of immunizations. As discussed above, in Malawi, only 3% of uncircumcised adult males who

received a voucher for a free circumcision at the local clinic underwent the surgery (Chinkhumba et al., 2014). This could be because of nonmonetary costs associated with take-up, such as time costs (e.g., walking 60 min to reach the facility where the free service is available), hassle costs (having to fill out complicated paperwork to receive a subsidy), and cultural barriers (for males thinking of getting circumcised as a means of reducing vulnerability to HIV). Some of these nonmonetary costs can be experimentally varied, in particular, distance and convenience. Other, most notably cultural psychic “costs,” cannot be directly experimented on, but their relative importance can sometimes be backed out by experimenting with financial incentives, the idea being that if a relatively small financial incentive increases adoption, then cultural barriers cannot be too important.

Back to the example from Udaipur, India, Banerjee et al. (2010) run an experiment to test the hypothesis that the reliability of the supply of free services may actually be at fault. The premise for this hypothesis is the observation made by Banerjee et al. (2004) (and discussed earlier in the chapter) that public facilities in charge of providing free immunization are characterized by very high absenteeism: spot checks conducted over a year suggested that 45% of the health staff were absent from their health posts, typically leading the health post to close, on any given workday. Because there was no predictable pattern to this absenteeism, obtaining all five shots included in the basic immunization package could require twice as many attempted visits to the public health facility.

In the experiment, some villages served as controls and other were randomly selected to receive a reliable, well-advertised “immunization camp.” The researchers found that the adding a reliable camp boosted full immunization rates from 6% to almost 18%. Impressive as the tripling of take-up in the immunization camps experiment may be, at only 18% fully immunized, the take-up rate remained among the lowest in the world. Could that be due to cultural barriers? In the experiment, a third group of villages was randomly selected to receive incentives for parents, in addition to the immunization camps. Specifically, parents were given a kilogram of lentils per immunization, and a set of plates for a child fully immunized. The incentive treatment increased immunization rates from 18% to 39%, which suggest that cultural barriers may not be decisive since they can be overcome with a fairly small handout. So what is the main barrier? We discuss the potential interpretations of this and other incentives experiment in the next section.

Thornton (2008) conducted a field experiment in rural Malawi that randomized the distance that individuals had to travel in order to obtain results of an HIV test, as well as whether they received a financial incentive to seek their results. This field experiment took place in 2005, at a time that preceded the introduction of rapid HIV tests, thus people had to make two visits to the testing center in order to learn their status—a first visit to get their blood drawn and a second visit a few weeks later to fetch their result. At the time Thornton conducted her experiment, the prevailing conventional wisdom in HIV prevention circles was that demand for knowing one’s status was very low due to the high psychic costs of learning one was HIV positive: since there was no access to antiretroviral

therapy (ARV) in Malawi at the time, learning one's positive status was akin to a death sentence. If psychic costs were indeed high, then providing a financial incentive and reducing the time costs of fetching one's results to a minimal level would likely have only a small effect on the demand for test results.

[Thornton \(2008\)](#) found the exact opposite: providing a financial incentive increased the likelihood that individuals sought their HIV test results from 35% to 78%. Reducing the distance that one had to travel to get results also increased the share of individuals seeking their results. In the absence of any incentives, those living within 1.5 km from the center where results could be picked up were 6.4 percentage points more likely to seek their HIV results than those living more than 1.5 km away. These large impacts thus teach us two things: that distance matters, i.e., time costs are not to be neglected; and that psychic costs were, on the other hand, much less important than believed at the time.

In their study of the demand for contraceptives in Zambia, the vouchers that [Ashraf et al. \(2014\)](#) distributed granted appointments for ordinary family planning services (provided routinely at government clinics), but with a guarantee that the wait would be less than 1 h, and that the modern contraceptive method of their choice would be available. Take up of the voucher was high (47%), indicating, as in [Banerjee et al. \(2010\)](#), that unreliable supply and its associated substantial time costs may be important barriers to the take-up of services offered at health facilities.

The fact that nonmonetary costs matter can sometimes be used to improve targeting. Recall the results from the pricing experiments discussed in [Section 5.1](#), which on the whole suggested that in environments where people face serious liquidity constraints, as in most of the developing world, price is not a particularly good screening mechanism. It fails at allocating scarce products to those who have higher returns to these products, as many people with a high valuation for a product may not be able to afford it. On the other hand, under free distribution, the product may be wasted on people with a low valuation for the product, and for products where there is a high share of low-valuation individuals in the population, this may be very costly.

Imagine that a provider delivers a year's supply of water-treatment product to a household, and the household members learn within a few days that they hate the taste of chlorinated water and stop using the product. In such cases, where households need to learn their own valuation, imposing some nonmonetary cost that households have to pay to access the free product may be efficient. This is what the literature refers to as an "ordeal mechanism." The provider may, for example, require that those who want a year's supply go to a store to redeem coupons every month for 12 months. A field experiment conducted in Western Kenya in 2007–08 suggests that such a micro-ordeal can help target free products only to those who will use them. [Dupas et al. \(2016\)](#) provided households with the opportunity to obtain enough free samples of chlorine solution for the treatment of drinking water for a whole year, but varied the effort required to obtain

the samples. Households randomly allocated to treatment arm 1 received a free supply of chlorine delivered directly at their home, while households randomly allocated to treatment arm 2 were given 12 coupons which could be redeemed for chlorine at a local shop over the course of a year. The researchers compare chlorine usage across arms and find no difference in rates of usage, during the year that followed the distribution, between those who were required to redeem coupons, compared to those who were given chlorine directly. They estimate that under reasonable assumptions regarding distribution costs, the results imply a significant increase in cost-effectiveness, with no negative impact on usage, of imposing a nonmonetary price on the acquisition of a health good such as chlorine, which is not valued equally by all households.

[Ma et al. \(2014\)](#) perform a similar micro-ordeal experiment when estimating the demand for and usage of eyeglasses for schoolchildren in China. Recall from the discussion in [Section 5.2](#) that this study, performed in 2012–13, had found that usage of free eyeglasses (delivered at school) among primary school children in China was lower than 50%. This suggests that eyeglasses are more like chlorine than bed nets—the cost of using them outweighs the perceived benefits for a large share of the population. This is therefore an area where screening the unlikely nonusers prior to free distribution could be highly cost-saving. [Ma et al. \(2014\)](#) adopt the voucher approach of [Dupas et al. \(2016\)](#) as a “micro-ordeal”—parents had to redeem a voucher for free eyeglasses for their child at an optical store in the county’s center, about 25 km away on average—and find that it successfully screens out households whose child would not use the eye glasses were they delivered for free at school. When interacted with the information training program discussed earlier, however, the ordeal mechanisms appears counterproductive: it reduced the impact of the information. This result is interesting and suggests additional research on micro-ordeal is necessary to better understand when they are desirable. Indeed, it could be that the micro-ordeal mechanism undermined the information treatment in [Ma et al. \(2014\)](#)—if recipients interpreted the voucher scheme as a lower level of endorsement of the product on the part of the program implementers, compared to the endorsement that comes with free distribution at school.

## 5.6 Incentive experiments

Earlier we briefly mentioned two experiments that included incentives as one of their treatment arms: the Udaipur immunization study ([Banerjee et al., 2010](#)) and the HIV testing study ([Thornton, 2008](#)). In both cases, small financial incentives were provided to encourage take-up of a specific health behavior: immunizing children, and learning one’s HIV status, respectively. In both cases, the small financial incentives had a large impact on take-up. [Banerjee et al. \(2010\)](#) interpret this as possible evidence of present bias: the natural tendency to delay an action that is slightly costly today even if it has

high payoffs in the future, a tendency which can be overcome if the incentive is sufficient to transform the cost into a positive *immediate* benefit.

While these two studies consider incentives rewarding a specific behavior (a specific input in the health production function), more recent studies have looked at the impact of output-based incentives, namely, incentives paid based on achieving a certain health outcome. [de Walque et al. \(2012\)](#) estimated the impact of offering varying amounts of a cash incentive to remain STI-free among adults aged 18–30 years old in Tanzania. They randomly assigned 2409 individuals to one of three groups: (1) a “high-value” conditional cash transfer (CCT) group that received \$20 for testing negative for curable STIs; (2) a “low-value” CCT group that received \$10 for testing negative for curable STIs; and (3) a comparison group that received no transfer. STI tests were conducted for all groups every 4 months over 1 year. Over the course of the first year, the number of people who tested positive for STI infection significantly decreased among people who received the high-value CCT, but no reduction was found for the group that received the low-value CCT. One downside of this study is that it was not powered to detect effects on HIV infection, given lower prevalence (relative to the treatable STIs).

In a similar study in Malawi, [Kohler and Thornton \(2012\)](#) randomly provided cash transfers of random amounts to 1307 participants, which ranged from no cash to approximately US\$ 16, conditional on maintaining one’s HIV negative status for 1 year. Researchers conducted interviews with participants throughout the year to collect data on sexual behavior. The promise of financial incentives of any amount had no effect on subsequent self-reported sexual behavior or HIV status. However, receiving cash after the final round of HIV testing had significant effects on respondents’ self-reported behavior: specifically, men were 9 percentage points more likely to engage in riskier sex, and women were 6.7 percentage points less likely to do so. As in the case of the incentivized immunization experiment discussed above, these results suggest that money given in the present may have stronger effects on behavior than rewards in the future—but sometimes for worse, as in this case.

## 5.7 Psychology experiments

A host of studies in developed country contexts have explored how insights from psychology might be harnessed to increase adoption of preventive health behaviors. Indeed, in contexts where information, access and affordability are (largely) nonissues, the role of human behavior becomes the primary driver of health outcomes, and numerous surveys document that individuals have difficulty keeping up with their intentions when it comes to many health choices, including maintaining a regular exercise regimen, dieting, preventive screenings, etc. This has been coined the “want/should” conflict, pervasive in many domains besides health ([Bitterly et al., 2016](#)).

One insight from psychology concerns the role of *planning* in solving the “want/should” conflict. It has been argued that forming precise plans can reduce forgetfulness and procrastination by linking intended behaviors with a concrete future moment and course of action. Consistent with this, randomized experiments in the United States have found that prompting people to form plans about where and when (not only the precise day, but even the precise time) they will complete an intended health behavior significantly increases the likelihood that they follow through with it, be it vaccination (Milkman et al., 2011) or preventive screening (Milkman et al., 2013).

Related prompts have been attempted with a cross-cutting design in some of the pricing experiments described above. In particular, Dugas (2009) evaluated the impact of having individuals verbally commit to purchase the bed net product: a randomly selected half of all the households were asked about their intention to redeem the voucher for the bed net, and over 92% said “yes.” They were then asked to state who would sleep under it once they had bought it, and also asked to estimate how many days they would need before they were in a position to purchase the net. At the end of this discussion, the enumerator asked: “Ok, so this means that if I stop by the shop on *[date after the date provided by respondent]* I will find your deemed voucher there? Can you promise? Do we have a deal?” This verbal commitment intervention had no impact on actual redemption behavior. In particular, it made no difference to the slope of the demand curve.

This could be because people were victims of the so-called “planning fallacy” (Buehler et al., 2002): it seems they underestimated the time it would take them to save enough to afford the bed net. Kremer and Miguel (2007) similarly had earlier found no impact of an explicit verbal commitment intervention embedded as a treatment arm within their deworming take-up experiment in Kenya.

In some cases, the key issue may not be a want/should conflict but rather individual undecidedness. An extensive literature in psychology and marketing suggests that decision-making can be affected by frames or cues that do not add information about a product, but can be effective at persuading undecided individuals to invest in it. The aforementioned experiment in Dugas (2009) evaluated the effects of framing, besides those of price and planning. At the time they received their first voucher, households in her experiment were exposed to a randomly assigned marketing message. The “health framing” group emphasized the morbidity and mortality due to malaria which could be avoided by using the bed net. The “financial framing” group emphasized the financial gains households would realize (from averting medical costs and loss of daily income) if they could prevent malaria. A third group received no marketing message. Neither of the two framing options (health or financial) had any impact on bed net take up. Combined with the finding that usage is quasiuniversal among free bed net recipients, this finding suggests that the price sensitivity of bed net demand is largely due to liquidity constraints rather than undecidedness.

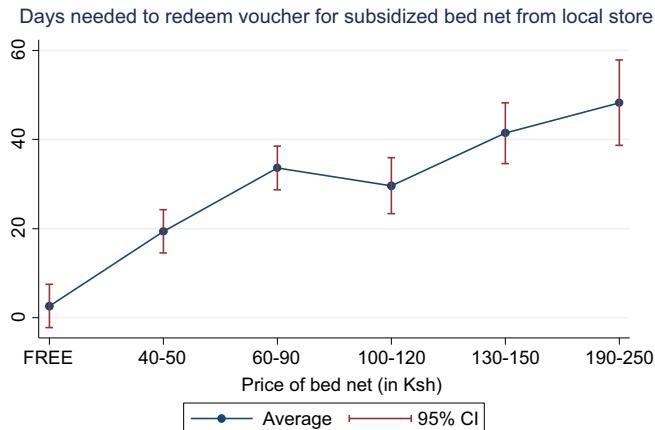
A third important factor in health behavior is how much attention people have to devote to it. Adherence to daily health treatment regimens is often imperfect, and “I forgot” is a common answer to the question of why a patient did not take his or her treatment as planned. Against this back-drop, [Pop-Eleches et al. \(2011\)](#) estimate the effect of reminders on adherence to antiretroviral treatment among AIDS patients in Kenya. 430 adult patients who had initiated antiretroviral therapy (ART) within 3 months were assigned to either a comparison group that received no reminders, or to one of four groups that received either: (1) a short message sent daily, (2) a long message (that included words of encouragement) sent daily, (3) the short message sent weekly, or (4) the long message sent weekly. Treatment adherence was monitored through bottles with a medication-event-monitoring cap.

The results indicate that 53% of those in the weekly reminder groups achieved treatment adherence greater than 90%, compared to just 40% in the comparison group. Yet daily reminders had no impact on individual ART adherence, suggesting that patients respond less to a frequently repeated stimulus. The type of message also had no impact on drug adherence. There is a need for future research to understand why daily reminders in this context were ignored (in particular, were they so frequent to get treated as “spam” and not even opened?) while weekly reminders had a substantial effect. It would also be useful to explore the role of different types of framing beyond those explored here.

## **5.8 Taking stock: how important is the role of present bias in explaining observed preventive health behaviors?**

In a chapter entitled “Improving Health in Developing Countries: Evidence from Randomized Evaluations” written for the *Handbook of Health Economics*, [Kremer and Glennerster \(2011\)](#), henceforth GK) review a subset of the experiments described in Sections 5.1–5.8 (namely, those available at the time of writing) and conclude that they “provide considerable support for a present bias model.” Specifically, they argue that present bias is useful in explaining why small prices or convenience barriers can dramatically reduce take-up of cost-effective approaches to prevention and nonacute care. In this section we revisit their conclusions in light of the accumulating evidence in this area, including the latest work.

The main argument underlying GK’s conjecture that present bias is an important factor is the following. In the standard human capital investment model they propose, in which individuals invest in a given health behavior or product if the expected discounted private benefit exceeds the cost, there is no reason to expect that a large proportion of the population would switch into using a given preventive product by the change from a low positive price to a zero price, unless there are large disutility costs that almost exactly offset the benefits. As they write, “the odds that this would occur for multiple products in multiple settings are particularly low” unless the discount rate is very high for a large group of people, but they rightfully argue that this would not be consistent with the



**Figure 3** Delays in redemption of vouchers for subsidized nets, by price. (Courtesy of Dugas, P., 2014a. *Short-run subsidies and long-run adoption of new health products: evidence from a field experiment*. *Econometrica* 82 (1), 197–228.)

fact that many people in the studies considered also invest in their children's education, which likely generates only modest and very delayed returns.

In contrast, in a model with present bias, a small change of price away from zero can have a large impact as the small cost is given a lot of weight against the discounted future benefit. Present bias could also explain why small incentives, such as the lentils provided in the [Banerjee et al. \(2010\)](#) immunization experiment in rural India, could lead to a sharp increase in take-up: the authors of that study write that “Providing the lentils helps to overcome procrastination because the lentils make the occasion [the visit to the immunization center] a small ‘plus’ rather than a small ‘minus’. Thus, in the case of preventive care, small barriers might turn out to have large implications.”

While we agree with GK's theoretical discussion, our interpretation of the evidence from the growing body of pricing experiments discussed in [Section 5.1](#) differs somewhat from theirs. In particular, as shown in [Fig. 2](#), for most products there does not appear to be anything special about zero, although there are exceptions (namely, for deworming and chlorine in Kenya, as noted above).<sup>5</sup>

Although the pricing experiments may not provide clear evidence that present bias (in terms of consumption) is a primary barrier to adoption of health products, one could also look for empirical evidence on the importance of present bias (and related self-control issues) in situations where monetary costs are unimportant but effort or hassle costs might

<sup>5</sup> Note that many experiments did not include free distribution (price of zero), which does limit interpretation somewhat. However, for most of the studies the smallest price point considered is fairly low and demand at that price very substantial. E.g., the fact that 93% of Kenyan households purchased slippers at the price point of 5 Ksh (USD 0.08) in [Meredith et al. \(2013\)](#), suggests that there is not a discontinuity in demand at 0.

cause procrastination. TIOLI experiments carried out with coupons (vouchers) redeemable at a prespecified location (in contrast to the door-to-door visit experiments) provide a way to examine that issue. In particular, we would expect present-bias to lead many people to procrastinate in redeeming their voucher until it is about to expire. Yet the data on the timing of voucher redemption shown in Fig. A1 of [Dupas \(2014a\)](#), reproduced here as [Fig. 3](#), suggests that most recipients of free bed net vouchers redeemed them immediately, suggesting any present bias in effort (if present) was not sufficiently large to overwhelm the positive expected health benefits. Similarly, none of the other studies that used vouchers ([Meredith et al., 2013](#); [Dupas et al., 2015, 2016](#); [Peletz et al., 2016](#)) report a bunching in redemption around the expiry date. This is a strike against the present-bias interpretation.

It is also potentially useful to revisit the present bias interpretation of the immunization incentives discussed in [Banerjee et al. \(2010\)](#). In that experiment, 78% of parents took their child to obtain the first injection without the incentive, and 75% took their child to obtain a second injection. The impact of the lentils incentive was thus mainly to “reduce the number of children dropping out after three injections.” This suggests that the hassle cost (immunization was free, but it took some time and effort to go to the immunization camp which was up to 5 km away and where there could be some waiting time; moreover, the child might develop a fever afterwards) are not large enough to deter take-up of the first couple of injections. What is an issue is the fact that completion of the whole course (5 rounds of injections) is low (only 25% without incentives compared to 40% with incentives). The drop-off after three injections could thus be due to a lack of understanding that the later rounds of injections (the booster shots) are actually important. Observing the impact of immunization on health may be difficult in noisy environments were illness episodes are very common and due to many different causes—the difference in the incidence of health shocks among immunized children versus nonimmunized children over the first year of life may not be perceptible with the naked eye, and this may contribute to low perceived benefits of immunization. In such a context, the incentives could possibly have been perceived as a signal that completing the full course is in fact important, the same way transfers to parents labeled as “for education” raised the perceived returns to education in Morocco ([Benhassine et al., 2015](#)). Or it could be simply be that if the perceived benefits are very low, and the cost of going to the immunization camp is low enough that people are nearly indifferent, then even small incentives are sufficient to shift their decision.

Possibly the most clear-cut evidence in favor of the present bias model would come from empirical evidence of an individual taste for commitment. The definition provided in [Bryan et al. \(2010\)](#) is that a commitment device is “any arrangement, entered into by an individual, with the aim of making it easier to fulfill his or her own future plans.” In the next [Section 5.8](#), we discuss three studies of commitment products for health behavior change from developing country contexts, all three finding that a subset of

the population does take-up commitment schemes against future selves—though whether these work at affecting health outcomes is unclear. This emerging evidence from developing country contexts is reminiscent of a number of studies from developed country contexts, finding strong evidence of a demand for commitment to healthier lifestyles. For example, [DellaVigna and Malmendier \(2006\)](#) show that individuals purchase a flat-fee membership to the gym as an (often unsuccessful) commitment device to increase future attendance: in their data, 89% of new gym users sign up for a monthly contract, even though given their actual attendance frequency, 80% would have saved money choosing the pay-per-visit contract, holding constant the number of visits. The role of present bias in health decision-making may thus grow in developing countries as the burden of disease tilts towards noncommunicable diseases (NCDs) like diabetes and hypertension, which have seen a substantial rise in prevalence (or at least, detection) in countries as varied as India, China, and Mexico in recent years, as well as in parts of Sub-Saharan Africa. Daily disease management, in particular lifestyle changes in terms of diet and exercise, are key to decreasing the rate of complications and avert early mortality from NCDs. Thus, to the extent that self-control and procrastination issues are particularly important for diet and exercising, there will be a growing need for health experiments in low-income countries testing behavioral change interventions that tackle time inconsistency in preferences.

## 5.9 Commitment experiments

A growing body of empirical evidence from many domains suggests that people seek commitment devices to help themselves follow through on their plans. Commitment devices can be “hard” or “soft.” An experiment involving a hard commitment device for health is the study of the Committed Action to Reduce and End Smoking (“CARES”) program, a voluntary commitment savings program to stop smoking, designed and tested by [Giné et al. \(2010\)](#) in collaboration with a bank in the Philippines. The basic design of the program was as follows: a smoker could open a bank account and deposit a self-selected amount of his own money that would be forfeited unless he passes a urine test indicating smoking cessation after 6 months. Regular smokers were recruited for the study off the street. All subjects received an informational pamphlet on the dangers of smoking, and a tip sheet on how to quit.

The commitment contract was taken up by 11% of smokers offered the account, and on average participants made deposits every two weeks and had a balance of 585 (US\$ 11) pesos after 6 months, some 535 pesos more than the minimum balance and approximately 6 months’ worth of cigarette spending. Individuals who were *offered* a CARES contract were 3.3–5.8 percentage points more likely to pass a urine test (negative for nicotine) after 6 months than those in the comparison group, and were 3.4–5.7 percentage points more likely to pass it after 12 months, a substantial effect considering

the well-known difficulties of quitting and the fact that only 8.9% to 14.7% of comparison individuals passed the test. This represents an over 35% increase in the likelihood of smoking cessation compared to baseline. Treatment on the treated effect estimates imply a 30–65% quit rate for this population relative to the control group. However, the overall welfare impact of offering the commitment contract program is unclear. A large proportion (66%) of smokers who voluntarily committed to CARES ended up failing to quit, and thus lost the money they had deposited.

[Schilbach \(2015\)](#) estimates the demand for commitment for *sobriety* among cycle-rickshaw drivers in the city of Chennai, India. The sample for the study was restricted to rickshaw drivers with a significant average daily consumption of hard liquor at baseline. At some point in the study, the drivers in the sample were given the choice between either (a) a guaranteed payment of 150 Rupees or (b) a guaranteed payment of 90 Rupees and an additional 30 Rupees conditional on them passing a breathalyzer test. The payment scheme under option (b) is *de facto* an incentive to be sober. Since the maximum possible value of the payment under the incentive scheme is lower than under option (a), everyone should choose option (a), unless people *value* the incentive embedded in option (b). [Schilbach \(2015\)](#) finds that about 30% of those given the choice chose the incentive payment—thus losing at minimum 30 Rupees (\$0.50) for sure (and since many of them failed to pass the test, the actual loss was greater than that). When the choice was between option (a') with 120 guarantee versus the same option (b) as above, the share taking up the incentive scheme was around 50%. Here again, a large share of those who took up the commitment device failed to pass the breathalyzer test and thus ended up earning less money than if they had chosen option (a'). The effect of the sobriety incentives on actual drinking behavior was modest—the incentives reduced day-time drinking significantly but not overall drinking, due to substitution towards night time drinking.

“Softer” commitment devices, which do not require that individuals put their own money on the line (be it savings or potential earnings), were studied by [Dupas and Robinson \(2013\)](#), who examine the determinants of health savings. They find a large demand for a commitment savings product that earmarks savings for health emergencies, as well as for a simple product (namely, a lockable box) that people can use to store their health savings. Both products have a large and significant impact on health investments, with the former decreasing the risk that an illness goes untreated and the second increasing spending on preventive health, compared to a control group that was also primed to save for health but not provided any specific tools to facilitate saving. In contrast, very little was saved in a product that locked up savings for preventative health with no possibility to access the savings in times of emergency, suggesting that some degree of flexibility is essential in product design.

## 6. SUPPLY OF HEALTH CARE

Most of the experiments described above in [Section 5](#) examine determinants of the demand for *preventive health*. This is because demand for acute care is quite high, as discussed previously in [Dugas \(2011a\)](#) and [Kremer and Glennerster \(2011\)](#).

A growing body of evidence also documents important gaps in both access and quality in the delivery of health services in developing countries, especially for the poor ([World Bank, 2004](#); see [Das and Hammer, 2014](#) for a review). Major issues identified to date concern: absenteeism among public health providers ([Chaudhury et al., 2006](#); [Banerjee et al., 2004](#); [Banerjee et al., 2008](#)); limited knowledge and training, as well as an important “know-do” gap among health professionals (see [Das et al., 2008a](#) for a review); limited availability of diagnostic testing, leading to high rates of inappropriate treatment ([Banerjee et al., 2004](#); [Cohen et al., 2015](#)); and drug quality problems ([Bennett and Yin, 2014](#); [Nayyar et al., 2012](#)). Another common concern is that of corruption among health providers, although quantitative evidence on this is limited, and the evidence to date is perhaps less pessimistic than anticipated ([Dizon-Ross et al., 2016](#)). The majority of experimental economics research on these issues has focused on testing the effectiveness of interventions aimed at improving quality by changing how providers are monitored and/or incentivized. Only a handful of experimental studies have seriously attempted to tackle the problems of limited diagnostic availability and drug quality.

The healthcare market in most countries in the developing world is comprised of government facilities with trained professionals, adjacent to a myriad of loosely regulated informal providers, from quack doctors to drug shops staffed by individuals with no formal pharmaceutical training but from whom medical advice is regularly sought. Incentives for healthcare providers at public facilities to come to work, or to perform well while at work, are generally seen as very weak ([World Bank, 2004](#); [Das and Hammer, 2014](#)). In contrast, private providers face at least some market discipline, but their ability to perform may be limited by inadequate medical training, and given the traditional information asymmetry in the health sector, patients’ ability to avoid low quality informal providers may also be limited. In this set-up, there are two ways to improve the quality of the health services that the majority of the poor have access to. The first is to better improve incentives for trained providers in the public sector. The second is to improve the quality of informal providers through better training.

Below we first review the set of studies that document the types of problems observed in the health care market, before turning to experiments that aimed to provide solutions to some of these problems.

### 6.1 Experimental audit studies

The last 15 years have seen a great deal of innovation in how empirical audit studies can be used to measure the quality of health care provision. An important innovation is the

use of a type of audit called the standardized patient (SP) method, which has long been used for training purposes but has now been adapted to allow quality measurements in “business as usual” conditions, and is considered the gold standard in assessing the quality of medical care delivered in outpatient settings. Standardized patients are people from a local community who are trained as actors to present a given “case” (symptoms) to a health provider. They are trained to find ways to refuse examinations that would reveal that they are in fact lying about some symptoms (e.g., temperature checks if they report having a fever). After the visit with the provider, they then record the details of their interaction, and in particular, any questions asked by the provider, any examinations done, and the diagnosis that was pronounced (if any). This method and other measurement tools are well described and discussed in the World Bank manual *Are You Being Served? New Tools for Measuring Service Delivery* ([World Bank, 2008](#)), to which we refer the reader for more details. In what follows, we describe some of the key insights generated by this new type of measurement tool.

[Das et al. \(2012\)](#) use the SP method with a representative sample of providers to document provider quality in both rural and urban areas of Madhya Pradesh in 2008–09. The study revealed a number of issues, including that: the majority of health care providers are not medically trained; overall quality is poor in both urban and rural settings and across all types of providers, with the average provider spending less than 4 min with a patient, and low rates of diagnosis (let alone correct diagnosis), low rates of correct treatment, and massive levels of overtreatment. Similar results were found in a replication of the study among 48 providers in a Chinese province ([Sylvia et al., 2015](#)).

Following on this work, [Das et al. \(2016\)](#) use the SP method to compare the quality of care provided by a given provider depending on whether the patient visited him in his private practice or at the public facility. Indeed, even though it is illegal, around 61% of public providers with a medical degree moonlight (i.e., run a separate private practice on the side). In private clinics, providers are paid on a fee-for-service basis by their customers. They can also earn a profit from selling medication. In contrast, as public servants doctors are paid a fixed salary, and the drugs they prescribe are supposed to be provided for free at the public clinic. Patient demand for treatment appears very high in the context studied, however, so there are possibly demand-driven incentives to overprescribe among both public and private providers.

There are 71 providers in this [Das et al. \(2016\)](#) study that were visited by standardized patients in both their public and private clinics, and analysis generates three main findings: (1) providers exert greater effort when the SP visits their private rather than public practice; (2) the likelihood that the “correct treatment” is prescribed is higher in the private clinic, and the likelihood that a palliative treatment (which relieves pain but does not solve the problem) is prescribed is lower; and (3) the likelihood of incorrect or even harmful treatment is identical across the two settings. The total number of drugs prescribed is also identical across the two settings. The authors conclude that the incentives

generated through customer accountability in the fee-for-service private market lead to greater provider effort and the higher rate of correct treatment being prescribed, in stark contrast to the solely administrative (and *ipso facto* nonexistent) accountability in the public sector. Interestingly, this gain in provider performance does not come at the expense of increased overtreatment: the likelihood of incorrect treatment does not increase in the private clinics relative to the public setting, but rather as private providers exert greater effort in examining the patient (and thus acquire tighter priors on the likely illness), they are more likely to prescribe the correct drug.

A country where medical overtreatment is considered a particularly important issue is China, where antibiotic abuse in particular has been well-documented. Using simulated (standardized) patients, Currie et al. (2011) provide evidence that a large share of this overprescription behavior is initiated by the physician: while none of their simulated patients required antibiotics given the symptoms they experienced, 62% were prescribed antibiotics. The question then becomes what incentives physicians face that leads them to overprescribe. There are two leading hypotheses: first, they could be responding to what they think is the patients' wish, or second, they could be responding to financial incentives. The setup in China is such that doctors do indeed have strong financial incentives to overprescribe drugs: while wage employees at public hospitals, they get paid a bonus that represents typically over a third of their salary, and this bonus is in part based on sales that they generate at the attached hospital pharmacy.

Several pieces of evidence suggest that financial incentives are indeed a key factor behind drug overprescription. First, Currie et al. (2011) find that physicians tend to prescribe more expensive rather than less expensive antibiotics. Second, Currie et al. (2011) carried out a randomization across the simulated patients regarding whether the patients would express knowledge of appropriate antibiotic use to the physician (during the visit). They find that such demonstrated knowledge by the patient reduces the incidence of unwarranted antibiotic prescription by 25 percentage points (40%). While this could be consistent with physicians having vastly mistaken beliefs about the extent to which consumers demand antibiotics, the authors argue it more likely reflects the fact that once physicians realize the patient will not buy the prescribed antibiotic in any case (due to their medical knowledge), the financial incentive to overprescribe is gone. To further test this point, in a follow-up audit study experiment Currie et al. (2014) randomized (within physician) whether the simulated patients: (1) said nothing special (the control group); (2) directly asked the doctor for an antibiotic prescription; (3) asked for a prescription (not specifically antibiotics) but indicated that he/she would buy any drugs prescribed in another pharmacy, thereby eliminating the financial incentive for the physician; (4) asked specifically for antibiotics and indicated that he/she would buy any drugs prescribed elsewhere. As in the previous study, overprescription was very high, with 55% of physicians prescribing antibiotics when the patient said nothing (the control). This rate increased to a staggering 85% when patients specifically requested

antibiotics, but only if the physician expected the prescription to be filled in the hospital pharmacy. If the patient indicated that he/she would purchase the drugs elsewhere, antibiotics were prescribed in only 14% of the cases, even when antibiotics had been specifically requested by the patient. Among patients who did not request antibiotics but indicated that they would buy any type of drug elsewhere, antibiotics were prescribed at a rate of 10%, statistically indistinguishable from the 14% if antibiotics had been requested. These results strongly indicate that high rates of antibiotic prescription are not mainly driven by patient demand or provider ignorance, but rather by providers' misaligned financial incentives.

These audit studies above have focused on physician behavior. [Dizon-Ross et al. \(2016\)](#) audit the performance of other health workers, primarily nurses and midwives, as they are asked to implement a bed net distribution program targeted to pregnant women, an increasingly common scheme in Sub-Saharan Africa. They conduct these audits in Kenya and Uganda (where the distribution scheme is a government program), and in Ghana, where the distribution scheme is sponsored by a nongovernmental organization. In all three cases, they find relatively satisfactory performance levels among providers, with the vast majority of eligible beneficiaries receiving the subsidized bed net as intended. They also measure whether health workers respond to bribe attempts from ineligibles by sending “mystery clients”—undercover enumerators posing as ineligible individuals trying to obtain a bed net. They find that a very small minority of mystery clients were successful at obtaining a bed net from prenatal centers, suggesting that in the context of an easily observable targeting rule (only pregnant woman are eligible) health workers comply with it, the same way Chinese physicians stop overprescribing when they know their patient are knowledgeable about appropriate treatment ([Currie et al., 2011](#)).

Another important dimension of health care quality that can be measured through experimental audit studies concerns drugs, in particular, the prevalence of counterfeits. [Bennett and Yin \(2014\)](#) sent mystery shoppers to small drug stores in Hyderabad, India. The mystery shoppers bought two common off-patent antibiotics, which were then analyzed in a lab. They found that 6% of the samples fell below pharmacopeia standards. Among so-called local (cheaper) brands, the share counterfeited was as high as 22%. In a comparable study, [Bjorkman-Nyqvist et al. \(2014b\)](#) tested the quality of antimalarials (ACTs) in Uganda, and found that around 30% were spurious/fake.

## 6.2 Monitoring experiments

To better align the incentives of trained providers with the overall objective of improving health outcomes, the interventions studied to date can be grouped into four primary types: district-level contracting, input-based incentives (e.g., nurses are paid a bonus, or avert a fine, if their absenteeism is sufficiently low), output-based incentives (e.g.,

providers are paid a bonus if health outcomes in their community are sufficiently high), and decentralization (giving greater monitoring power to local communities). We discuss each in turn below.

First, however, we note that many of the same types of interventions have also been assessed in the education sector in low income countries, and a number of findings in that literature are likely to be relevant to health. For brevity, we do not survey the education experiments here, however, and instead refer the interested reader to the chapter by Muralidharan in this volume. We also note that there have been many related provider performance experiments in the field of public health, although they are also not our focus here.

### **6.2.1 District-level contracting**

In what is arguably among the most innovative experiments ever carried out in the health sector, in 1999 Cambodia launched a brand new approach to solving the low quality problem: it contracted out management of government health services to private (international) nonprofit organizations (NGOs) in some districts. In eight districts randomly selected out of 12 districts involved in this at-scale “pilot,” NGOs could bid for the government contract, which was paid for through increased public-health expenditures. Contracts were ultimately signed in only five of the eight districts, as in two districts no bids were received, and in a third district the bid was too expensive and was not chosen by the government.

In the five contracted districts, contractors were responsible for health services at all levels, from district hospitals down to remote health posts. Performance on eight key (targeted) service-delivery indicators (most of them related to maternal and child health) was measured, and the contract was renewed yearly based on these outcomes. The idea behind this district-level contracting scheme is that it can strengthen incentives for government workers while reducing potentially harmful incentives associated with private fee-for-service provision, such as the incentive to overprescribe antibiotics discussed above.

[Bhushan et al. \(2007\)](#) study whether the contracting in Cambodia indeed improved health performance by using the randomized assignment to contracting as an instrument for a district actually having a private contractor. They use baseline survey data from 1997 and follow-up survey data from 2003 in the analysis, and find meaningful improvements in health-care service delivery, especially on the targeted indicators, such as receipt of vitamin A by children under 5, which increased by 21 percentage points, and receipt of antenatal care by pregnant women, which increased by 33 percentage points. These improvements also did not come at the expense of nontargeted health services, which experienced no decline. Instead, the gains on the targeted indicators appear to have come about through improvements in management quality: absenteeism of providers and stock outs of drugs and other equipment fell in districts where the NGOs were in

charge. This increase in public facility service reliability in turn increased demand for these services, with residents increasing the number of visits made to public facilities and reducing their visits to (often expensive) informal providers, such as traditional healers.

### **6.2.2 Top-down, input-based incentives**

Absenteeism among public health providers has been shown to be a very significant concern in many parts of the developing world ([Chaudhury et al., 2006](#)). Since provider presence is obviously a necessary if not sufficient condition for health care services to function, a number of experiments in the health sector (alongside with a number of experiments in the education sector, plagued by the same absenteeism problem) consider the effectiveness of programs incentivizing providers present. We discuss them in this section.

[Banerjee et al. \(2008\)](#) evaluated an incentives program for Assistant Nurse Midwives (ANM) at Primary Health Subcenters in Udaipur District in the Indian state of Rajasthan. The program was implemented collaboratively by a nonprofit organization and the state and local health administrations, with the goal of improving ANM's attendance at rural subcenters. Indeed earlier research had established that due to pervasive absenteeism among ANMs, health centers were closed 56% of the time during regular business hours ([Banerjee et al., 2004](#)). The program tested consisted of monitoring ANM attendance and “punishing” absenteeism: ANMs absent for more than 50% of the time on monitored days would have their pay reduced proportional to the number of absences recorded that month, and ANMs absent for more than 50% of the time on monitored days for a 2nd month would be suspended from government service. The program was implemented in 49 randomly selected subcenters. In those centers, the ANM was required to stamp a register secured to the wall of the subcenter three times a day: once at 9 a.m., once between 11 a.m. and 1 p.m., and once at 3 p.m., using a tamper-proof time/date-stamping machine. Researchers then measured the impact of the program on ANM performance through random unannounced visits to the 49 “treatment” subcenters and 51 control subcenters.

The results of this intervention are mixed. In the short-run, the incentive scheme was highly successful, doubling attendance, from around 30–60%. The program was not popular with nurses, however, who complained heavily to the local health administration about the pay deductions. The share of “missed stamps” due to either an (intentionally) broken time clock or excused absence increased considerably over time, and at 16 months after program inception, the absence rates were comparable between treatment and control centers. What the researchers take away from these mixed results is that, on the one hand, nurses are responsive to properly administered incentives, but on the other hand, incentive systems can be very difficult to properly administer, due to the perennial question of “who monitors the monitor?” Ultimately, the decision to

monitor and incentivize public sector employees is a political one, and there may be a variety of political economy explanations for why these programs are opposed by either public employee unions, or the public at large.

A similar experiment took place a few years later with primary health care center staff in Karnataka ([Dhaliwal and Hanna, 2013](#)). Instead of an NGO, the program was designed and implemented by the National Rural Health Mission (NRHM) of Karnataka, the lead state department for the delivery of health services. Instead of stamps, the monitoring system relied on fingerprints taken at the beginning and end of each day, and instead of proportional pay deductions, the penalty was a loss of paid vacation days, although in practice the penalty was rarely imposed. These researchers found that the health staff monitoring system increased attendance among medical staff by 18%, but not among doctors. They also find a large, 26% decrease in the incidence of low birth weight, confirming that provider attendance is potentially a critical input in the health production function, although effects on a range of other health outcomes are mixed. The mechanism through which birth weight was affected appears not to have been through an increase in prenatal care attendance, but rather an increase in the likelihood that prenatal clients received iron folic acid tablets. Once again, as in the earlier Indian health monitoring and incentive experiment, the program did not appear to be politically sustainable.

### ***6.2.3 Top-down, output-based incentives***

While provider attendance may be the first step, it may not always lead to improved population health outcomes. For this reason, more recent monitoring experiments have based the rewards on outcomes rather than on inputs such as provider attendance. Basing incentives on actual health outcomes is difficult, however, for reasons discussed in [Miller and Babiarz \(2014\)](#), in particular, the fact that provider behavior is only one of many factors that determine health outcomes, and that health outcomes can be difficult and expensive to measure. Given this, outcomes over which performance-pay contracts are written tend to relate to the utilization or coverage of specific easily observable health services, e.g., the share of children who are immunized, the share of pregnant women seeking prenatal care, the share of deliveries that take place at the facility, etc. The potential downside of contracting over such specific and narrow indicators is that providers may devote too much effort to those, at the expense of activities related to noncontracted indicators which may be just as important for the production of health but simply harder to measure (in a version of the multitasking problem).

At the time of writing, we are aware of two economics field experiments that have directly tested the impact of performance-based incentives in the health sector, one in Indonesia and one in Rwanda. In both cases, the incentives were at the group level (not at the individual provider level), and the performance mattered for the total budget available to the providers, rather than for their own personal gain. (As far as we know,

performance-based incentives for individual health workers have not yet been studied using field experimental methods.)

The Indonesia experiment estimated the effect of incentivized community-based block grants that aimed to improve both health and education (Olken et al., 2014). The program, known locally as Generasi, provided villages with annual block grants of \$8000 to \$14,000, and villages were encouraged to use the funds to make progress on 12 prespecified maternal and child health indicators, including prenatal visits, delivery by trained midwives, childhood immunizations, and child growth monitoring. For the experiment, which was conducted jointly with the Government of Indonesia, 264 subdistricts were randomized into either control or to one of the two versions of the Generasi program: the “incentivized” version with a pay-for-performance component, or the otherwise identical, “nonincentivized” version without pay-for-performance incentives. In the first year of the program, villages in all groups received program funds based on their size and demography. In the 2nd year, the allocation rule stayed the same for the nonincentivized villages, but for the incentivized villages 20% of the funds were distributed based on the village’s performance on the 12 indicators during year 1. Impacts were measured over the 2 years.

The pay-for-performance incentives led to an increase in the labor hours of midwives, the major providers of maternal and child health services in the area. Likely as a result, the targeted maternal and child indicators were somewhat higher in incentivized villages than in nonincentivized villages, but the overall effect was quite modest, with a gap of just 0.04 standard deviations on average. The main impacts were on the number of prenatal visits (an 8.2% increase) and regular monthly weight checks for children under five (+4.5%). The effect of the incentives varied with the baseline levels of service delivery, however, and effects were stronger in the poorer provinces not on Java. Interestingly, no detrimental effects of the incentive scheme were detected on nontargeted health indicators at least to the extent they could be measured.

The Rwanda experiment was conducted in partnership with the government as it launched a national pay-for-performance scheme to supplement primary health center budgets (Basinga et al., 2011). As a pilot, the program was supposed to be launched first in 80 facilities from eight randomly chosen districts, with eight districts (86 facilities) assigned to a comparison group. Under the program, facilities received payments as a function of their performance on 14 maternal and child health-care output indicators, including many of those used in the Indonesia study. Performance was assessed as follows: facilities in the program had to submit monthly activity reports which were then audited against the facility’s records. The specific payment amounts differed for each service, between US\$ 0.09 for an initial prenatal visit and US\$ 4.59 for an institutional childbirth delivery. Facilities in the control group received funding as a function of their size and the demographic characteristic of their catchment area.

Unfortunately, the experimental design was compromised somewhat before the start of the study due to a change in district boundaries that required that the research team switch treatment and control status for eight districts. The final design is thus more of a quasiexperimental design, and the authors use a difference-in-difference estimation strategy to study impacts. They estimate large positive impacts on some of the targeted indicators. In particular, the incentives led to a 23% increase in the number of institutional deliveries, a 56% increase in the number of preventive care visits by children aged 0–2 years, and a massive 132% increase in the number of preventive care visits by older children. They also found a 0.16 standard deviations increase in prenatal quality as measured by compliance with Rwandan prenatal care clinical practice guidelines, but no change in the quantity of prenatal care sought or in rates of full immunization among children.

One of the mechanisms underlying the estimated effects appear to be an increase in health provider productivity. The researchers measured productivity as “the gap between provider knowledge and actual practice of appropriate prenatal care clinical procedures” ([Gertler and Vermeersch, 2013](#)). This gap appears substantial in the control group: while providers know 63% of the appropriate clinical protocols for prenatal care on average (based on correct answers when asked), they appear to only deliver about 45% of the appropriate protocols. This 18 percentage point gap was reduced by 4 percentage points in the incentivized facilities. The gap is much larger at baseline among providers who have better knowledge and skills, and the impact of the pay-for-performance incentives is larger for this subgroup.

#### **6.2.4 Bottom-up: beneficiary oversight**

While monitoring coupled with incentives—whether carrots or sticks, and whether input or output based—can be successful at improving provider performance, as demonstrated in the studied surveyed above, these programs can often be costly to implement. The monitoring costs may become prohibitive in remote areas, and as discussed earlier, they often generate the problem of “who monitors the monitor?,” as well as a political backlash among the staff who are not subjected to the monitoring. For this reason, an obvious alternative would be to make the monitor the person who is the direct beneficiary of the gains to be had, namely, the patient herself. In other words, citizens, as *clients* of healthcare providers, have a direct interest in seeing their performance improve and this could translate into a willingness to expend some effort (or cost) carrying out monitoring. The difficulty here is that of monitoring ability: how do patients know if their doctor is actually making the right diagnosis or correct prescription? Health care is one of many domains—in the popular consciousness, along with auto repair, plumbing, etc.—where clients often find it difficult to evaluate the performance of the informed “expert” providing the service. For this reason, the impact of increasing monitoring by beneficiaries by itself may be limited.

The existing experimental evidence to date on this issue comes from two experiments conducted in Uganda. In the first experiment conducted in nine Ugandan districts, [Bjorkman and Svensson \(2009\)](#) partnered with an NGO that focused on increasing the local accountability of health providers. The experiment was conducted with 50 communities (with one facility each), with 25 treatment and 25 control communities. In the treatment communities, the NGO first created “report cards” on the quality of services at the health facility, based on information generated through facility audits as well as household interviews conducted by the researchers. A unique report card was established for each facility, and it contained (1) information on key areas subject to improvement, including utilization, access, absenteeism, and patient–clinician interaction; and (2) comparisons vis-à-vis other local health facilities, and with the national standard for primary health care provision. The report cards were written in the local vernacular languages and included graphics to help communicate the key points to nonliterate residents.

The NGO then facilitated three sets of meetings: a provider staff meeting, a community meeting, and an interface meeting. The staff meeting was fairly short and consisted of sharing and discussing the content of the report cards. The community meeting gathered around 150 community members (the NGO made sure all stakeholders were represented, including the young, elderly, and women). Participants were asked to critically review the quality of the health services available to them locally in an open discussion, and through this process the NGO disseminated the information on the report cards. Participants were then encouraged to identify concrete steps the local providers could take to improve quality, as well as actions community members could take to monitor the providers taking those steps. The discussion and proposed solutions were summarized at the end of the meeting in an action plan. The content of the action plans differed across communities, but the researchers note that high rates of absenteeism, long waiting times, weak attention by the health staff, and differential treatment across residents were common to many of the 25 communities in the treatment group. The interface meeting encouraged community members and health workers to discuss patient rights and provider responsibilities. At the end of the interface meeting, the community and the facility staff reached an agreement on the way forward. This shared action plan was called a “community contract” and it spelled out concrete steps for the provider to take and specific ways that the community members would monitor them. The NGO came back after 6 months to conduct two additional meetings, a community and an interface meeting, to discuss the progress to date and fine-tune the action plan.

To estimate impacts, the researchers conducted surveys 1 year after the first set of three meetings had taken place. They surveyed households (the same households surveyed prior to the intervention and whose information was used for the report cards), as well as the health staff. They also collected administrative records from the facilities and performed visual checks. They find large positive impacts on both the quality of care

and on health outcomes. A year after the first round of meetings, health facilities in treatment communities had taken significant steps to reduce waiting times, in particular through the introduction of numbered waiting cards (20% of the treatment facilities had them, compared to only 4% of control) and a 13% reduction in absenteeism, leading to a reduction in waiting times of 12 min on average. This is despite the fact that utilization of general outpatient services in the local public facility was 20% higher in the treatment communities, with households shifting away from traditional healers and self-treatment. In particular, immunizations increased for all age groups, especially newborns, and prenatal care attendance also increased, contributing to a 0.14 z-score increase in infant weight and a remarkable 33% reduction in under-5 mortality. Encouragingly, these large effects persisted over time: 4 years after the initial intervention, researchers went back to collect new data, and found that the treatment communities still had significantly higher rates of health care utilization, better adherence to clinical guidelines by providers, and better health outcomes, including reduced child mortality and increased weight-for-age and height-for-age for children ([Bjorkman-Nyqvist et al., 2014c](#)).

The results of this first experiment suggest that beneficiary oversight can work, but it is notable that in this study beneficiaries were given information that they could act on, as well as, even if implicitly, information on how to stay informed (the report cards gave them concrete items to look for when monitoring). Does beneficiary control work in the absence of this fairly costly provision of information? The second experiment, by the same researchers in the same Ugandan context, suggests that the answer is probably no ([Bjorkman-Nyqvist et al., 2014c](#)). In new communities, the researchers designed an intervention that mimicked everything in the first experiment except for the report cards. The new intervention is called “participation only,” in contrast to the “participation and information” treatment of the first experiment, and in this setting the staff and community meetings started without any quantitative (or even qualitative) information being provided by the NGO.

The researchers found no significant impact of this “participation only” intervention on health provision or outcomes. The authors conclude that information provision is key, and theorize that it enables clients to better distinguish between health workers’ actual effort versus factors that also matter for outcomes but are outside the health workers’ control. This information thus makes effective monitoring possible, since the client knows what to focus his monitoring efforts on, and this beneficiary monitoring is what then leads to improved health worker performance.

### 6.3 Improving the quality of informal providers

Although the experiments described above suggest that increasing the accountability of public health providers can improve their productivity, the extent to which this will

affect population health outcomes depends on the “market share” of these public providers. While services such as prenatal care are rarely provided outside regulated facilities, curative primary care is commonly provided by informal private sector providers with at best minimal medical training. The large role played by poorly trained “quack” doctors has been well documented in India, in particular (see [Das and Hammer, 2014](#), and references therein). In Sub-Saharan Africa, it is common for households to procure medication through retail sector drug shops without consulting public providers first ([Cohen et al., 2015](#)). In such contexts, while increasing widespread availability of quality public care may be the goal in the long run, improving the quality of the care and/or medical advice provided by informal providers may be essential in the short run. In this section, we discuss two recent economics experiments, one in India and one in Uganda, of programs designed to improve the quality of services and products available outside the formal sector.

In West Bengal, India, [Banerjee et al. \(2015b,c\)](#) estimate the impact of offering training to existing informal providers (IPs). The training program included 72 sessions of 2 h, spread over 9 months, and was taught by certified medical doctors, but no training certificate was issued upon completion. The training course covered multiple illnesses, and emphasis was placed on basic medical conditions, triage, and avoidance of harmful practices, accompanied by frequent patient simulations. Informal providers could continue operating their clinics throughout the training since the training demanded only 4 h of time per week. Nevertheless, take-up of the training was not universal: out of 360 providers initially asked whether they were interested in the training program, 304 (84.5%) initially signed up. Half of those 304 were then randomly assigned to start the training immediately (the treatment group). Of these 152 treatment IPs, 20 (or 13%) quit the program within three sessions, bringing take-up to just above 70%. The attendance rate over the training period was then 64% among those IPs that took up the program.

To measure impacts, the researchers used unannounced standardized patients. As described above, this method is considered the “gold standard” in care quality measurement because it does not suffer from observation and recall bias, and because it generates estimates of both the quality of the diagnosis (since illnesses are prespecified in the study design) and of the treatment prescribed conditional on the diagnosis. In this West Bengal experiment, SP’s were trained to depict symptoms of either angina, asthma, or dysentery in a child asleep at home. These three conditions require different dimensions of care (angina requires referral, asthma requires identification of a chronic lung disease, and dysentery requires the provision of ORS) which were all supposed to be impacted by the training. The data collection through SPs started 3 months after the completion of training, and SPs were completed for 267 of the 304 providers in the study sample. Additional data was collected through clinical observations.

The researchers found a significant, positive impact of the training on the quality of care. Being assigned to the training group improved case-specific checklist adherence by 4.2 percentage points (on a base of 27.3% in the control group) and the likelihood of correct treatment by 7.8 percentage points (from 52% to 59.8%). Prescription of antibiotics (unnecessary in all cases) remained unchanged at very high levels (close to 50%), though interestingly, such unnecessary or even harmful practice is even more common in the public sector. Patients may thus mistakenly expect to be prescribed medicines in all cases, leading trained informal providers to continue prescribing them in order to keep their clients satisfied. Overall though, the results of this West Bengal experiment suggest that training existing IPs can improve the quality of care for rural populations with little access to care from fully qualified providers in either the public or private sector.

When existing providers are absent, or unwilling to go for training, an alternative is to encourage entry of new higher quality providers. Bjorkman-Nyqvist et al. (2014a) evaluate the impact of market-based community health care program led by two NGOs, BRAC and Living Goods, in Uganda. In treatment villages (107 villages randomly chosen out of 204), the NGOs recruited and trained one woman per village to be a “Community Health Promoter” as well as an incentivized sales agent, conducting home visits to not only educate households on essential health behaviors but also sell preventive and curative health products at 20–30% below prevailing retail prices, with the woman earning a margin on product sales. Data collected from households 3 years after the rollout of the intervention in treatment villages suggests that the introduction of these trained informal providers considerably increased care seeking and resulted in a large 25% reduction in under-5 mortality.

One potential channel through which the community health promoter program improved outcomes could have been by influencing other actors to improve the quality of services and products that they provided or sold. In a companion paper, the researchers document that this may have been the case with respect to drug quality (Bjorkman-Nyqvist et al., 2014b). The quality of the antimalarials that community health promoters were allowed to sell was strictly monitored by the NGOs, and the authors argue that this reduced the likelihood that antimalarials sold at drug shops in the treatment villages were counterfeits, through a pro-competitive effect. Exploiting the randomized assignment of the program across villages, and using data on drug quality from a subset of villages, they estimate that the introduction of the community health promoter in the village led to an increase in the share of authentic artemisinin-based antimalarials sold by incumbent drug shops of 11–13 percentage points, corresponding to a roughly 50% decrease in the share of fake drugs.

The take-away from this experiment is not entirely clear-cut, however. Its results suggest that subsidizing high-quality products in the private sector improves health outcomes, but the cost of such subsidization could be prohibitive, especially if one includes the cost that the implementing NGOs had to incur for quality control

purposes. Moreover, whether this subsidization requires that the promoters are paid through their sales rather than a fixed salary is unclear. The NGOs running the programs consider the implicit piece-rate pay system to be a critical feature of their model (which they call “entrepreneurial”), but the experiment was not designed to estimate the role of the incentive-pay in observed impacts. Introducing these sorts of incentive pay schemes may be more difficult in other settings, especially in the public sector health context.

If the market is sufficiently large, subsidization may not be needed, however: pharmacy chains are common in developed countries and increasingly common in developing ones, and thanks to economies of scale, it seems that they can often guarantee better quality at low prices. [Bennett and Yin \(2014\)](#) discuss how chains can improve quality by purchasing in bulk from trusted manufacturers, establishing independent distribution networks, employing licensed pharmacists, and advertising to raise consumer awareness. Studying the impact of the entry of chain pharmacies in many markets of Hyderabad, India, they find that the entry of chain pharmacies selling only high quality products led to market-wide impacts on quality just like those observed in the Ugandan experiment.

## 7. CONCLUSION

We have surveyed the large and growing literature using field experiments to study issues of health and health care access, usage and impacts in low income countries. There has been a veritable explosion of research in this area: 20 years ago, there was almost no experimental research in development economics, and we are in a position today where a (very long) chapter such as this one is only able to cover a small share of the published research literature in any depth. As we have argued, research progress has been particularly pronounced in the study of the demand for health products and services, the quality of health care, and in certain aspects of the question of health impacts. However, there remain a number of glaring gaps in the literature, and promising areas for future investigation, and we briefly survey these topics in this concluding section.

One of the most important areas of inquiry is how current adult health status affects contemporaneous labor productivity. However, this is also an area where there remain relatively few well-designed field experiments (with [Thomas et al., 2003, 2006](#) being notable exceptions). One possible explanation is that relatively few field experiments on health in low income settings have taken place with private sector commercial partners, those who would be most likely to have access to a large pool of workers who might serve as participants in such a health intervention study. (As discussed above, most experiments on health in development economics have been conducted with government or NGO partners.) This remains a topic of great intellectual and public policy importance, and one where further research could have high returns.

A related limitation of existing research is the relative lack of work studying the long-run impacts of earlier health investments. While research evidence is beginning to accumulate in this area—including the long-term follow-ups to the well-known INCAP (Hoddinott et al., 2008) and Kenya deworming (Baird et al., 2015) cases—few studies are successfully able to combine field experimental research designs with long-term longitudinal data collection with high tracking rates. Such studies require long time horizons and prospective planning, not to mention large research budgets, and few have been successfully carried out. Putting in place the data collection plans to follow-up the participants in the large number of recent health experiments in development economics going forward could have a large research payoff. In the absence of such studies, most of our understanding about the long-run impacts of child health status comes from studies that rely on natural experiments (such as weather variation, as in the Maccini and Yang, 2009 study from Indonesia), or variation in exposure to conflict or other large-scale political shocks (as in the work of Alderman et al., 2006a,b, Bundervoet et al., 2009; Leon, 2012; among many other recent papers), but in these cases the interpretation of resulting impacts is arguably less transparent than in a well-designed field experiment.

Another broad area where further research would be useful is in the area of mental health, as well as the interaction between psychology and economics. While a growing number of studies, including many surveyed above, include measures of mental health, depression, anxiety and wellbeing as outcomes of particular health interventions (including in the WISE project discussed in Thomas et al., 2003, 2006 study, the KLPS data used in Baird et al., 2015, among others), relatively few studies by economists explore the effects of mental health interventions on economic or other life outcomes, or study the heterogeneity in the impacts of other interventions (e.g., information, cash drop) by underlying psychological status, such as stress level. This is despite the fact that there is mounting evidence that many mental health problems are widespread in both high and low income societies, and that there is a strong correlation in the cross-section between mental health and socio-economic outcomes (see the evidence in Das et al., 2008b and the references therein). The growing emphasis on psychological issues in economics overall and in the study of poverty, and in particular how particular psychological or neurophysiological processes affect the decision-making of the poor (for instance, see recent work by Mani et al., 2013; Haushofer and Fehr, 2014), suggests that this is an area that is also ripe for further intellectual exploration.

Beyond deepening our understanding of links between health outcomes (along various dimensions) and economic outcomes, it would also be valuable to direct more research energy to understand the large-scale health systems reforms that have occurred in many low income countries during the past decade. For instance, many less developed countries have expanded government supported health insurance programs—including in Ghana, India, Indonesia, Mexico, and Rwanda, among many other countries—but relatively little academic health economics research has examined the performance,

structure, and incentives produced by these reforms, or the resulting behavioral responses by households and individuals. This is in stark contrast to the health economics literature focusing on wealthy countries like the US, where much research energy has focused precisely on broader institutional and organizational issues in the health sector. While there is already some important work in this area (e.g., [Gertler and Vermeersch, 2013](#); [Miller and Babiarz, 2014](#)), development economists working on health could learn much from the public finance and industrial organization economists working on related health systems issues in the US, Europe, and other high income regions.

Over the past two decades, as field experiments have become an increasingly common tool used by applied researchers in development economics and health (and other fields of economics), we have often looked to the methods and approaches used in clinical trials as a model on which to base our own work. That was certainly the case with much of the early experimentation in economics (as discussed in [Duflo et al., 2007](#)), and has continued with the increasing use of study preregistration and pre-analysis plans over the past few years ([Miguel et al., 2014](#)). There has been extensive and productive learning across disciplines. However, the increasing sophistication of the methods used to study health issues in low income countries, and the innovations in research design and measurements described above, now make it more likely, in our view, that considerable learning will flow in the opposite direction, from economics and other social science disciplines back into medical research. This appears especially likely given the growing awareness in global health research that the study of health policies, systems and individual behavior cannot be approached in the same way as traditional medical treatment efficacy trials (for a discussion of the rise of “implementation science” in health research, see [Madon et al., 2007](#), and [Mwisongo et al., 2011](#)). It is thus our hope that the discussion in this chapter will not only be of interest to development economists already working on health in low income settings, but may also prove useful to scholars in other fields or disciplines who are engaged with these issues.

## REFERENCES

- [Adhvaryu, A., Kala, N., Nyshadham, A., 2014a. Management and Shocks to Worker Productivity: Evidence From Air Pollution Exposure in an Indian Garment Factory. University of Michigan \(unpublished working paper\).](#)
- [Adhvaryu, A., Kala, N., Nyshadham, A., 2014b. The Light and the Heat: Productivity Co-benefits of Energy-Saving Technology. University of Michigan \(unpublished working paper\).](#)
- [AEA RCT Registry, 2013. American Economics Association’s Registry for Randomized Controlled Trials. <http://www.socialscienceregistry.org>.](#)
- [Ahuja, A., Kremer, M., Zwane, A.P., 2010. Providing safe water: evidence from randomized evaluations. \*Annu. Rev. Econ.\* 2, 237–256.](#)
- [Aiken, A., Davey, C., Hargreaves, J., Hayes, R., 2015. Re-analysis of health and educational impacts of a school-based deworming program in western Kenya: a pure replication. \*Int. J. Epidemiol.\* <http://dx.doi.org/10.1093/ije/dyv127>.](#)

- Alderman, H., Hoddinott, J., Kinsey, B., 2006a. Long term consequences of early child malnutrition. *Oxf. Econ. Pap.* 58 (3), 450–474.
- Alderman, H., Konde-Lule, J., Sebuliba, I., Bundy, D., Hall, A., 2006b. Increased weight gain in preschool children due to mass albendazole treatment given during 'Child Health Days' in Uganda: a cluster randomized controlled trial. *Br. Med. J.* 333, 122–126.
- Alderman, H., 2007. Improving nutrition through community growth promotion: longitudinal study of nutrition and early child development program in Uganda. *World Dev.* 35 (8), 1376–1389.
- Ali, M., Emch, M., von Seidlein, L., Yunus, M., Sack, D.A., et al., 2005. Herd immunity conferred by killed oral cholera vaccines in Bangladesh: a reanalysis. *Lancet* 366, 44–49. [http://dx.doi.org/10.1016/S0140-6736\(05\)66550-6](http://dx.doi.org/10.1016/S0140-6736(05)66550-6).
- Ali, M., Emch, M., Yunus, M., Sack, D., Lopez, A.L., et al., 2008. Vaccine protection of Bangladeshi infants and young children against cholera: implications for vaccine deployment and person-to-person transmission. *Pediatr. Infect. Dis. J.* 27, 33–37. <http://dx.doi.org/10.1097/INF.0b013e318149dffd>.
- Ali, M., Sur, D., You, Y.A., Kanungo, S., Sah, B., et al., 2013. Herd protection by a bivalent-killed-whole-cell oral cholera vaccine in the slums of Kolkata, India. *Clin. Infect. Dis.* <http://dx.doi.org/10.1093/cid/cit009> pii:cit009.
- Almond, D., Currie, J., 2010. Human Capital Development Before Age Five. NBER. Working paper #15827.
- Amarante, V., Manacorda, M., Miguel, E., Vigorito, A., 2016. Do cash transfers improve birth outcomes? Evidence from matched vital statistics, program, and social security data. *Am. Econ. J. Econ. Policy* 8.
- Angelucci, M., Di Maro, V., 2015. Program Evaluation and Spillover Effects. University of Michigan (unpublished work paper).
- Arrow, K.J., 1963. Uncertainty and the welfare economics of medical care. *Am. Econ. Rev.* 53 (5), 941–973.
- Ashraf, N., Berry, J., Shapiro, J., 2010. Can higher prices stimulate product use? Evidence from a field experiment in Zambia. *Am. Econ. Rev.* 100 (5).
- Ashraf, N., Jack, B.K., Kamenica, E., 2013. Information and subsidies: complements or substitutes? *J. Econ. Behav. Organ.* 88, 133–139.
- Ashraf, N., Field, E., Lee, J., 2014. Household bargaining and excess fertility: an experimental study in Zambia. *Am. Econ. Rev.* 104 (7), 2210–2237.
- Attanasio, O., Fernandez, C., Fitzsimons, E.O.A., Grantham-McGregor, S.M., Meghir, C., Rubio-Codina, M., 2014. Using the infrastructure of a conditional cash transfer program to deliver a scalable integrated early child development program in Colombia: cluster randomized controlled trial. *Br. Med. J.* 349, g5785.
- Baird, S., Garfein, R.S., McIntosh, C., Ozler, B., 2012. Effect of a cash transfer programme for schooling on prevalence of HIV and herpes simplex type 2 in Malawi: a cluster randomised trial. *The Lancet* 379 (9823), 1320–1329.
- Baird, S., Hamory Hicks, J., Miguel, E., 2008. Tracking, Attrition and Data Quality in the Kenyan Life Panel Survey Round 1 (KLPS-1). University of California. CIDER Working paper.
- Baird, S., De Hoop, J., Özler, B., 2013. Income shocks and adolescent mental health. *J. Hum. Resour.* 48, 370–403.
- Baird, S., Bohren, A., McIntosh, C., Ozler, B., 2014. Designing Experiments to Measure Spillover Effects. George Washington University (unpublished working paper).
- Baird, S., Hicks, J.H., Kremer, M., Miguel, E., 2015. Worms at Work: Long-Run Impacts of a Child Health Investment. NBER. Working paper #21428.
- Banerjee, A., Barnhardt, S., Duflo, E., 2015a. Movies, margins and marketing: encouraging the adoption of iron fortified salt. In: Wise, D.A. (Ed.), *Insights in the Economics of Aging*, NBER Book Series. University of Chicago Press (Forthcoming).
- Banerjee, A., Chowdhury, A., Das, J., Hussam, R., 2015b. The Impact of Training Informal Providers on Clinical Practice in West Bengal, India: A Randomized Controlled Trial. Working paper.
- Banerjee, A., Deaton, A., Duflo, E., 2004. Health care delivery in rural Rajasthan. *Econ. Polit. Wkly.* 39 (9), 944–949.
- Banerjee, A.V., Duflo, E., Glennerster, R., 2008. Putting a band-aid on a corpse: incentives for nurses in the Indian public health care system. *J. Eur. Econ. Assoc.* 6 (2–3), 487–500.

- Banerjee, A., Duflo, E., Glennerster, R., Kothari, D., 2010. Improving immunization coverage in rural India: a clustered randomized controlled evaluation of immunization campaigns with and without incentives. *BMJ* 340, c2220.
- Banerjee, A., Duflo, E., Goldberg, N., Karlan, D., Osei, R., Parienté, W., Shapiro, J., Thuysbaert, B., Udry, C., May 15, 2015c. A multifaceted program causes lasting progress for the very poor: evidence from six countries. *Science* 348 (6236). <http://dx.doi.org/10.1126/science.1260799>.
- Banerjee, A., Karlan, D., Zinman, J., 2015d. Six randomized evaluations of microcredit: introduction and further steps. *Am. Econ. J. Appl. Econ.* 7 (1), 1–21.
- Basinga, P., Gertler, P.J., Binagwaho, A., Soucat, A.L.B., Sturdy, J., Vermeersch, C.M.J., 2011. Effect on maternal and child health services in Rwanda of payment to primary health-care providers for performance: an impact evaluation. *Lancet* 377, 1421–1428.
- Becker, G.M., Degroot, M.H., Marschak, J., 1964. Measuring utility by a single-response sequential method. *Syst. Res. Behav. Sci.* 9 (3), 226–232.
- Bennear, L., Tarozzi, A., Soumya, H.B., Pfaff, A., Ahmed, K.M., van Geen, L., 2013. Impact of a randomized controlled trial in arsenic risk communication on household water-source choices in Bangladesh. *J. Environ. Econ. Manag.* 65 (2), 225–240.
- Behrman, J.R., Calderon, M.C., Preston, S., Hoddinott, J., Martorell, R., Stein, A.D., November 2009. Nutritional supplementation of girls influences the growth of their children: prospective study in Guatemala. *Am. J. Clin. Nutr.* 90, 1372–1379.
- Benhassine, N., Devoto, F., Duflo, E., Dupas, P., Pouliquen, V., 2015. Turning a shove into a nudge: a ‘labeled cash transfer’ for education. *Am. Econ. J. Econ. Policy* 7 (3), 86–125.
- Benjamin-Chung, J., Colford, J., Berger, D., Arnold, B., Jimenez, V., Tran, D., Clark, A., Konagaya, E., Falcao, L., Abedin, J., Hubbard, A., Luby, S., Miguel, E., 2015. The Identification and Measurement of Health-Related Spillovers in Impact Evaluations: A Systematic Review (unpublished working paper).
- Bennet, D., Yin, W., 2014. The Market for High-Quality Medicine: Retail Chain Entry and Drug Quality in India. Mimeo, University of Chicago.
- Berry, J., Fischer, G., Guiteras, R., 2012. Eliciting and Utilizing Willingness to Pay: Evidence From Field Trials in Northern Ghana. Mimeo, London School of Economics.
- Bhattacharya, D., Dupas, P., Kanaya, S., 2013. Estimating the Impact of Means-tested Subsidies Under Treatment Externalities With Application to Anti-Malarial Bednets. University of Oxford (unpublished working paper).
- Bhusan, I., Bloom, E., Clingingsmith, D., Hong, R., King, E., Kremer, M., Loevinsohn, B., Schwartz, J.B., 2007. Contracting for Health: Evidence From Cambodia. Mimeo.
- Bitterly, T.B., Mislavsky, R., Dai, H., Milkman, K.L., 2015. Dueling with desire: a synthesis of past research on want/should conflict. In: Hoffman, W., Nordgren, L. (Eds.), *The Psychology of Desire*, pp. 244–264.
- Bjorkman-Nyqvist, M., Guariso, A., Svensson, J., Yanagizawa-Drott, D., 2014a. Evaluating the Impact of the Living Goods Entrepreneurial Model of Community Health Delivery in Uganda: A Cluster-randomized Controlled Trial. Working paper.
- Bjorkman, M., Svensson, J., 2009. Power to the people: evidence from a randomized field experiment on community-based monitoring in Uganda. *Q. J. Econ.* 124 (2), 735–769.
- Bjorkman-Nyqvist, M., Svensson, J., Yanagizawa-Drott, D., 2014b. Can Good Products Drive Out Bad? Evidence From Local Markets for Antimalarial Medicine in Uganda. Working paper.
- Bjorkman-Nyqvist, M., de Walque, D., Svensson, J., 2014c. Information Is Power: Experimental Evidence of the Long-run Impact of Community Based Monitoring. World Bank Policy Research. Paper Series No. 7015.
- Bleakley, H., 2007. Disease and development: evidence from hookworm eradication in the American South. *Q. J. Econ.* 122 (1), 73–117.
- Bleakley, H., 2010a. Malaria eradication in the Americas: a retrospective analysis of childhood exposure. *Am. Econ. J. Appl. Econ.* 2 (2), 1–45.
- Bleakley, H., 2010b. Health, human capital, and development. *Annu. Rev. Econ.* 2, 283–310.
- Bobonis, G., Miguel, E., Sharma, C.P., 2006. Iron deficiency, anemia and school participation. *J. Hum. Resour.* 41 (4), 692–721.

- Boozer, M., Philipson, T.J., 2000. The impact of public testing for immunodeficiency virus. *J. Hum. Resour.* 35 (3), 419–446.
- Bryan, G., Karlan, D., Nelson, S., 2010. Commitment devices. *Annu. Rev. Econ.* 2, 671–698.
- Buehler, R., Griffin, D., Ross, M., 2002. Inside the planning fallacy: the causes and consequences of optimistic time predictions. In: Gilovich, T., Griffin, D., Kahneman, D. (Eds.), *Heuristics and Biases: The Psychology of Intuitive Judgment*. Cambridge University Press, Cambridge, UK, pp. 250–270.
- Bundervoet, T.P.V., Akresh, R., 2009. Health and civil war in rural Burundi. *J. Hum. Resour.* 44 (2), 536–563.
- Bundy, D.A.P., 1988. Population ecology of intestinal helminth infections in human communities. *Philos. Trans. R. Soc. Lond. Ser. B* 321 (1207), 405–420.
- Bundy, D., Guyatt, J., 1996. Schools for health: focus on health, education, and the school-age child. *Parasitol. Today* 12, 1–16.
- Bundy, D.A.P., Chan, M.-S., Medley, G.F., Jamison, D., Savioli, L., 1998. Intestinal nematode infections. In: *Health Priorities and Burden of Disease Analysis: Methods and Applications From Global, National and Subnational Studies*. Harvard University Press for the World Health Organization and the World Bank.
- Casey, K., Glennerster, R., Miguel, E., 2012. Reshaping institutions: evidence on aid impacts using a pre-analysis plan. *Q. J. Econ.* 127 (4), 1755–1812.
- Chassang, S., i Miquel, G.P., Snowberg, E., 2012. Selective trials: a principal-agent approach to randomized controlled experiments. *Am. Econ. Rev.* 102 (4), 1279–1309.
- Chaudhury, N., Hammer, J., Kremer, M., Muralidharan, K., Rogers, F.H., 2006. Missing in action: teacher and health worker absence in developing countries. *J. Econ. Perspect.* 20 (1), 91–116.
- Chinkhumba, J., Godlonton, S., Thornton, R., 2014. The demand for medical male circumcision. *Am. Econ. J. Appl. Econ.* 6 (2), 152–177.
- Chong, A., Karlan, D., Gonzalez-Navarro, M., Valdivia, M., 2013. Effectiveness and Spillovers of Online Sex Education: Evidence From a Randomized Evaluation in Colombian Public Schools. Working paper.
- Chong, A., Cohen, I., Field, E., Nakasone, E., Torero, M., 2016. Iron deficiency and schooling attainment in Peru. *Am. Econ. J. Appl. Econ.* 8 (4), 222–255.
- Clark, S.E., et al., 2008. Effect of intermittent preventive treatment of malaria on health and education in schoolchildren: a cluster-randomised, double-blind, placebo-controlled trial. *Lancet* 372 (9633), 127–138.
- Clasen, T., Boisson, S., Routray, P., Torondel, B., Bell, M., Cumming, O., Ensink, J., et al., 2014. Effectiveness of a rural sanitation programme on diarrhoea, soil-transmitted helminth infection, and child malnutrition in Odisha, India: a cluster-randomised trial. *Lancet Glob. Health* 2, e645–e653.
- Coffman, L.C., Niederle, M., 2015. Pre-analysis plans have limited upside, especially where replications are feasible. *J. Econ. Perspect.* 29 (3), 81–98.
- Cohen, J., Dupas, P., 2010. Free Distribution or cost-sharing? Evidence from a randomized malaria experiment. *Q. J. Econ.* 125, 1–45.
- Cohen, J., Dupas, P., Schaner, S., 2015. Price subsidies, diagnostic tests, and targeting of malaria treatment. *Am. Econ. Rev.* 105 (2), 609–645.
- Cooper, E., Fitch, L., 1983. Pertussis: herd immunity and vaccination coverage in St. Lucia. *Lancet* 322, 1129–1132. [http://dx.doi.org/10.1016/S0140-6736\(83\)90637-2](http://dx.doi.org/10.1016/S0140-6736(83)90637-2).
- Cox, D.R., 1958. *Planning of Experiments*. Wiley, Oxford, England, 308 pp.
- Croke, K., 2014. The Long Run Effects of Early Childhood Deworming on Literacy and Numeracy: Evidence From Uganda. Harvard University (unpublished working paper).
- Currie, J., 2000. Child health in developed countries. In: Culyer, A.J., Newhouse, J.P. (Eds.), *Handbook of Health Economics*, vol. 1B. North-Holland, Amsterdam, pp. 1053–1090.
- Currie, J., Lin, W., Zhang, W., 2011. Patient knowledge and antibiotic abuse: evidence from an audit study in China. *J. Health Econ.* 30, 933–949.
- Currie, J., Lin, W., Meng, J.J., 2014. Addressing antibiotic abuse in China: an experimental audit study. *J. Dev. Econ.* 110, 39–51.
- Cutler, D., Miller, G., 2005a. Water, Water, Everywhere: Municipal Finance and Water Supply in American Cities. In: Glaeser, E., Goldin, C. (Eds.), *Corruption and Reform: Lessons from America's Economic History*. University of Chicago Press, Chicago, pp. 153–183.

- Cutler, D., Lleras-Muney, A., 2014. Education and health: insights from international comparisons. In: Culyer, A.J. (Ed.), *Encyclopedia of Health Economics*, vol. 1. Elsevier, San Diego, pp. 232–245.
- Cutler, D.M., Miller, G., 2005b. The role of public health improvements in health advances: the twentieth-century United States. *Demography* 42 (1), 1–22.
- Das, J., Hammer, J., Leonard, K., 2008a. The quality of medical advice in low-income countries. *J. Econ. Perspect.* 22 (2), 93–114.
- Das, J., Do, Q.T., Friedman, J., McKenzie, D., 2008b. Mental Health Patterns and Consequences: Results From Survey Data in Five Developing Countries. World Bank Policy Research. Working paper #4495.
- Das, J., Hammer, J., 2014. Quality of primary care in low-income countries: facts and economics. *Annu. Rev. Econ.* 6, 525–553.
- Das, J., Holla, A., Das, V., Mohanan, M., Tabak, D., Chan, B., 2012. The quality of medical care in clinics: evidence from a standardized patients study in a low-income setting. *Health Aff.* 31 (12), 2274–2784.
- Das, J., Holla, A., Mohpal, A., Muralidharan, K., 2016. Quality and accountability in health: audit evidence from primary care clinics in India. *Am. Econ. Rev.*, (forth coming).
- Davey, C., Aiken, A., Hayes, R., Hargreaves, J., 2015. Re-analysis of health and educational impacts of a school-based deworming programme in western Kenya: a statistical replication of a cluster quasirandomized stepped-wedge trial. *Int. J. Epidemiol.* <http://dx.doi.org/10.1093/ije/dyv128>.
- de Walque, D., Dow, W.H., Nathan, R., Abdul, R., Abilahi, F., Gong, E., Isdahl, Z., Jamison, J., Jullu, B., Krishnan, S., Majura, A., Miguel, E., Moncada, J., Mtenga, S., Mwanyangala, M.A., Packel, L., Schachter, J., Shirima, K., Medlin, C.A., 2012. Incentivising safe sex: a randomised trial of conditional cash transfers for HIV and sexually transmitted infection prevention in rural Tanzania. *BMJ* 2 (1), e000747.
- Deaton, A., 2013. *The Great Escape: Health, Wealth and the Origins of Inequality*. Princeton University Press.
- Delavande, A., Kohler, H.-P., 2012. The impact of HIV testing on subjective expectations and risky behavior in Malawi. *Demography* 49 (3), 1011–1036.
- DellaVigna, S., Malmendier, U., 2006. Paying not to go to the gym. *Am. Econ. Rev.* 96, 694–719.
- DeLong, J.B., Lang, K., 1992. Are all economic hypotheses false? *J. Polit. Econ.* 100 (6), 1257–1272.
- Devoto, F., Duflo, E., Dupas, P., Pariente, W., Pons, V., 2012. Happiness on tap: piped water adoption in urban Morocco. *Am. Econ. J. Econ. Policy* 4 (4), 68–99.
- Dewald, W.G., Thursby, J.G., Anderson, R.G., 1986. Replication in empirical economics: the journal of money, credit and banking project. *Am. Econ. Rev.* 76 (4), 587–603.
- Dhaliwal, I., Hanna, R., 2013. Deal with the Devil: The Successes and Limitations of Bureaucratic Reform in India. Working paper.
- Dickson, R., Awasthi, S., Williamson, P., Demellweek, C., Garner, P., 2000. Effect of treatment for intestinal helminth infection on growth and cognitive performance in children: systematic review of randomized trials. *Br. Med. J.* 320, 1697–1701.
- Dillon, A., Friedman, J., Serneels, P., 2014. Health Information, Treatment, and Worker Productivity. World Bank Policy Research. Working paper No. WPS 7120.
- Dizon-Ross, R., Dupas, P., Robinson, J., 2016. Governance and Effectiveness of Health Subsidies. NBER. Working paper 21324.
- Droitcour, J., Caspar, R.A., Hubbard, M.L., Ezzati, T.M., 1991. The item count technique as a method of indirect questioning: a review of its development and a case study application. In: Beimer, P.B., Groves, R.M., Lyberg, L.E., Mathiowetz, N.A., Sudman, S. (Eds.), *Measurement Errors in Surveys*. John Wiley & Sons, Inc, Hoboken, New Jersey, pp. 185–211.
- Duflo, E., Dupas, P., Kremer, M., 2015a. Education, HIV and early fertility: experimental evidence from Kenya. *Am. Econ. Rev.* 105 (9), 2257–2297.
- Duflo, E., Dupas, P., Kremer, M., 2016. The (over?)Promise of Education: Experimental Evidence From Ghana. Mimeo, Stanford University.
- Duflo, E., Glennerster, R., Kremer, M., 2007. Using randomization in development economics research: a toolkit. *Handb. Dev. Econ.* 4 (2007), 3895–3962.

- Duflo, E., Greenstone, M., Guiteras, R., Clasen, T., 2015b. Toilets Can Work: Short and Medium Run Health Impacts of Addressing Complementarities and Externalities in Water and Sanitation. Mimeo.
- Dupas, P., 2009. What matters (and what does not) in household's decision to invest in malaria prevention? *Am. Econ. Rev.* 99 (2), 224–230.
- Dupas, P., 2011a. Health behavior in developing countries. *Annu. Rev. Econ.* 3, 425–449.
- Dupas, P., 2011b. Do teenagers respond to HIV risk Information? Evidence from a field experiment in Kenya. *Am. Econ. J. Appl. Econ.* 3 (1), 1–36.
- Dupas, P., 2014a. Short-run subsidies and long-run adoption of new health products: evidence from a field experiment. *Econometrica* 82 (1), 197–228.
- Dupas, P., September 12, 2014b. Getting essential health products to their end users: subsidize, but how much? *Science* 345 (6202), 1279–1281.
- Dupas, P., Hoffmann, V., Kremer, M., Zwane, A., 2016. Targeting Health Subsidies Through a Nonprice Screening Mechanism: Evidence From a Randomized Controlled Trial in Rural Kenya. Mimeo, Stanford University.
- Dupas, P., Robinson, J., 2013. Why don't the poor save more? Evidence from health savings experiments. *Am. Econ. Rev.* 103 (4), 1138–1171.
- Dybvig, P.H., Spatt, C.S., 1983. Adoption externalities as public goods. *J. Public Econ.* 20, 231–247.
- Eble, A., Boone, P., Elbourne, D., 2015. On Minimizing the Risk of Bias in Randomized Controlled Trials in Economics. Brown University (unpublished working paper).
- Fafchamps, M., McKenzie, D., Quinn, S., Woodruff, C., 2014. Microenterprise growth and the flypaper effect: evidence from a randomized experiment in Ghana. *J. Dev. Econ.* 106, 211–226.
- Fine, P.E., 1993. Herd immunity: history, theory, practice. *Epidemiol. Rev.* 15, 265–302.
- Fink, G., Masiye, F., 2012. Assessing the impact of scaling-up bednet coverage through agricultural loan programmes: evidence from a cluster randomised controlled trial in Katete, Zambia. *Trans. R. Soc. Trop. Med. Hyg.* 106 (11), 660–667. <http://dx.doi.org/10.1016/j.trstmh.2012.07.01>.
- Fink, G., Masiye, F., July 2015. Health and agricultural productivity: evidence from Zambia. *J. Health Econ.* 42, 151–164.
- Finkelstein, A., et al., 2012. The Oregon health insurance experiment: evidence from the first year. *Q. J. Econ.* 127 (3), 1057–1106.
- Fitzsimons, E., Malde, B., Mesnard, A., Vera-Hernández, M., 2012. Household Responses to Information on Child Nutrition: Experimental Evidence From Malawi. Available: [http://papers.ssrn.com/sol3/papers.cfm?abstract\\_id=2034133](http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2034133).
- Forleo-Neto, E., Oliveira de, C.F., Maluf, E.M.C.P., Bataglin, C., Araujo, J.M.R., et al., 1999. Decreased point prevalence of *Haemophilus influenzae* type b (Hib) oropharyngeal colonization by mass immunization of Brazilian children less than 5 years old with hib polyribosylribitol phosphate polysaccharide—tetanus toxoid conjugate vaccine in combination with diphtheria-tetanus toxoids—pertussis vaccine. *J. Infect. Dis.* 180, 1153–1158. <http://dx.doi.org/10.1086/315018>.
- Fox, M., et al., 2004. The impact of HIV/AIDS on labour productivity in Kenya. *Trop. Med. Int. Health* 9 (3), 318–324.
- Franco, A., Malhotra, N., Simonovits, G., 2014. Publication bias in the social sciences: unlocking the file drawer. *Science* 345 (6203), 1502–1505.
- Galiani, S., Gertler, P., Schargrodsky, E., 2005. Water for life: the impact of the privatization of water services on child mortality. *J. Polit. Econ.* 113 (1), 83–120.
- Gertler, P., Vermeersch, C., 2013. Using Performance Incentives to Improve Health Outcomes. NBER. WP 19046.
- Gertler, P., et al., 2014. Labor market returns to an early childhood stimulation intervention in Jamaica. *Science* 344, 998–1001.
- Giné, X., Karlan, D., Zinman, J., 2010. Put your money where your butt is: a commitment contract for smoking cessation. *Am. Econ. J. Appl. Econ.* 2 (4), 213–235.
- Glewwe, P., Miguel, E., 2008. The impact of child health and nutrition on education in less developed countries. In: Schultz, T.P., Strauss, J. (Eds.), *Handbook of Development Economics*, vol. 4. Elsevier.

- Godlonton, S., Munthali, A., Thornton, R., 2016. Responding to risk: circumcision, information, and HIV prevention. *Rev. Econ. Stat.* 98 (2), 333–349.
- Godlonton, S., Thornton, R., 2012. Peer effects in learning HIV results. *J. Dev. Econ.* 97, 118–129.
- Gong, E., 2015. HIV testing and risky sexual behavior. *Econ. J.* 125, 32–60.
- Grossman, M., 1972. On the concept of health capital and the demand for health. *J. Polit. Econ.* 80, 223–255.
- Guiteras, R., Levinsohn, J., Mobarak, A.M., 2015. Encouraging sanitation investment in the developing world: a cluster-randomized trial. *Science* 348, 903–906.
- Habyarimana, J., Jack, W., 2011. Heckle and chide: results from a randomized road safety intervention in Kenya. *J. Public Econ.* 95 (2011), 1438–1446.
- Habyarimana, J., Jack, W., 2015. Results of a large-scale randomized behavior change intervention on road safety in Kenya. *Proc. Natl. Acad. Sci.* 4661–4670.
- Haushofer, J., Fehr, E., 2014. On the psychology of poverty. *Science* 344, 862–867.
- Haushofer, J., Shapiro, J., 2013. Household Response to Income Changes: Evidence From an Unconditional Cash Transfer Program in Kenya. Mimeo, Princeton University.
- Hicks, J.H., Kremer, M., Miguel, E., 2015. Commentary: deworming externalities and schooling impacts in Kenya: a comment on Aiken et al (2015) and Davey et al (2015). *Int. J. Epidemiol.* <http://dx.doi.org/10.1093/ije/dyv129>.
- Hoddinott, J., Maluccio, J.A., Behrman, J., Flores, R., Martorell, R., 2008. The impact of nutrition during early childhood on income, hours worked, and wages of Guatemalan adults. *Lancet* 371 (February), 411–416.
- Jayachandran, S., 2015. The roots of gender inequality in developing countries. *Annu. Rev. Econ.* 7 (2015), 63–88.
- Janison, J., Karlan, D., Raffler, P., May 2013. Mixed method evaluation of a passive mHealth sexual information texting service in Uganda. *Inf. Technol. Int. Dev.* 9 (3).
- Jensen, R., Lleras-Muney, A., 2012. Does staying in school (and not working) prevent teen drinking and smoking? *J. Health Econ.* 31 (4), 644–657.
- Joshi, S., Schultz, T.P., 2013. Family planning and women's and children's health: long-term consequences of an outreach program in Matlab, Bangladesh. *Demography* 50, 149–180.
- Karlan, D., Zinman, J., 2009. Observing unobservables: identifying information asymmetries with a consumer credit field experiment. *Econometrica* 77 (6), 1993–2008.
- Karlan, D., Zinman, J., 2011. List Randomization for Sensitive Behavior: An Application for Measuring Use of Loan Proceeds. NBER. Working paper #17475.
- Karlan, D., Fischer, G., McConnell, M., Raffler, P., 2014. To Charge or Not to Charge: Evidence From a Health Products Experiment in Uganda. Mimeo, Yale University.
- Kohler, H., Thornton, R., 2012. Conditional cash transfers and HIV/AIDS prevention: unconditionally promising? *World Bank. Econ. Rev.* 26 (2), 165–190.
- Kremer, M., Glennerster, R., 2011. Improving health in developing countries: evidence from randomized evaluations. In: Pauly, M.V., McGuire, T.G., Barros, P.P. (Eds.), *Handbook of Health Economics*, vol. 2, pp. 201–315.
- Kremer, M., Miguel, E., 2007. The illusion of sustainability. *Q. J. Econ.* 122 (3), 1007–1065.
- Kremer, M., Leino, J., Miguel, E., Zwane, A.P., 2011a. Spring cleaning: rural water impacts, valuation, and property rights institutions. *Q. J. Econ.* 126 (1), 145–205.
- Kremer, M., Miguel, E., Mullainathan, S., Null, C., Zwane, A., 2011b. Social Engineering: Evidence From a Suite of Take-up Experiments in Kenya. Mimeo, Emory University.
- Kvalsig, J.D., Cooppan, R.M., Connolly, K.J., 1991. The effects of parasite infections on cognitive processes in children. *Ann. Trop. Med. Parasitol.* 73, 501–506.
- LaBrie, J.W., Earleywine, M., 2000. Sexual risk behaviors and alcohol: higher base rates revealed using the unmatched-count technique. *J. Sex Res.* 37, 321–326.
- Laine, C., Horton, R., DeAngelis, C.D., Drazen, J.M., Frizelle, F.A., Godlee, F., Haug, C., et al., 2007. Clinical trial registration—looking back and moving ahead. *N. Engl. J. Med.* 356 (26), 2734–2736.
- Leamer, E.E., 1983. Let's take the con out of econometrics. *Am. Econ. Rev.* 73 (1), 31–43.

- Leon, G., 2012. Civil conflict and human capital accumulation: the long-term effects of political violence in Peru. *J. Hum. Resour.* 47 (4), 991–1022.
- Ma, X., Sylvia, S., Boswell, M., Rozelle, S., 2014. Ordeal Mechanisms and Training in the Provision of Subsidized Products in Developing Countries. Mimeo, Stanford University.
- Maccini, S., Yang, D., 2009. Under the weather: health, schooling, and economic consequences of early-life rainfall. *Am. Econ. Rev.* 99 (3), 1006–1026.
- Madajewicz, M., Pfa10, A., van Geen, A., Graziano, J., Hussein, I., Momotaj, H., Sylvi, R., Ahsan, H., 2007. Can information alone change behavior? Response to arsenic contamination of groundwater in Bangladesh. *J. Dev. Econ.* 84 (2), 731–754.
- Madon, T., Hofman, K.J., Kupfer, L., Glass, R.I., 2007. Implementation science. *Science* 318 (5857), 1728–1729.
- Maluccio, J.A., Hoddinott, J., Behrman, J.R., Martorell, R., Quisumbing, A., Stein, A.D., 2009. The impact of improving nutrition during early childhood on education among Guatemalan adults. *Econ. J.* 199 (537), 734–763.
- Mani, A., Mullainathan, S., Shafir, E., Zhao, J., 2013. Poverty impedes cognitive function. *Science* 341, 976–980.
- Martorell, R., Habicht, J.P., Rivera, J.A., 1995. History and design of the INCAP longitudinal study (1969–1977) and its follow-up (1988–89). *J. Nutr.* 125, 1027S–1041S.
- Meredith, et al., 2013. Keeping the doctor away: experimental evidence on investment in preventive health products. *J. Dev. Econ.* 105, 196–210.
- Miguel, E., Kremer, M., 2004. Worms: identifying impacts on education and health in the presence of treatment externalities. *Econometrica* 72.
- Miguel, E., Kremer, M., 2014. Worms: Identifying Impacts on Education and Health in the Presence of Treatment Externalities, Guide to Replication of Miguel and Kremer (2004). CEGA. Working paper #39.
- Miguel, E., Camerer, C., Casey, K., Cohen, J., Esterling, K.M., Gerber, A., Glennerster, R., et al., 2014. Promoting transparency in social science research. *Science* 343 (6166), 30–31. <http://dx.doi.org/10.1126/science.1245317>.
- Milkman, K.L., Beshears, J., Choi, J.J., Laibson, D., Madrian, B.C., 2011. Using implementation intentions prompts to enhance influenza vaccination rates. *Proc. Natl. Acad. Sci.* 108, 10415–10420.
- Milkman, K.L., Beshears, J., Choi, J.J., Laibson, D., Madrian, B.C., 2013. Planning prompts as a means of increasing preventive screening rates. *Prev. Med.* 56, 92–93.
- Miller, G., Babiarz, K.S., 2014. Pay-for-performance incentives in low- and middle-income country health programs. In: Culyer, A.J. (Ed.), *Encyclopedia of Health Economics*, vol. 2. Elsevier, San Diego, CA, pp. 457–483.
- Muralidharan, K., Sundararaman, V., 2011. Teacher performance pay: experimental evidence from India. *J. Polit. Econ.* 119 (1), 39–77.
- Muralidharan, K., Sundararaman, V., 2015. The aggregate effect of school choice: evidence from a two-stage experiment in India. *Q. J. Econ.* 130 (3), 1011–1066.
- Mwisingo, A., Wang, L., Madon, T., Owusu-Agyei, S., González Block, M.Á., 2011. Current and Foreseeable Themes in Implementation Research for Disease Control. Chapter 9, *Implementation Research for the Control of Infectious Diseases of Poverty*. World Health Organization, Geneva.
- Nayyar, G., Breman, J.G., Newton, P.N., Herrington, J., 2012. Poor-quality antimalarial drugs in southeast Asia and sub-Saharan Africa. *Lancet Infect. Dis.* 12 (6), 488–496.
- Neumark, D., 2001. The employment effects of minimum wages: evidence from a prespecified research design the employment effects of minimum wages. *Ind. Relat. J. Econ. Soc.* 40 (1), 121–144.
- Nokes, C., Grantham-McGregor, S., Sawyer, A., Cooper, E., Bundy, D., 1992. Parasitic helminth infection and cognitive function in school children. *Proc. Biol. Sci.* 247 (1319), 77–81.
- Nokes, C., van den Bosch, C., Bundy, D., 1998. The Effects of Iron Deficiency and Anemia on Mental and Motor Performance, Educational Achievement, and Behavior in Children: A Report of the International Nutritional Anemia Consultative Group. USAID, Washington, DC.
- Olken, B., Onishi, J., Wong, S., 2014. Should aid reward performance? Evidence from a field experiment on health and education in Indonesia. *Am. Econ. J. Appl. Econ.* 6 (4), 1–34.

- Olken, B., 2015. Promises and perils of pre-analysis plans. *J. Econ. Perspect.* 29 (3), 61–80.
- Ozier, O., 2014. Exploiting Externalities to Estimate the Long-Term Effects of Early Childhood Deworming. World Bank Policy Research Working paper #7052.
- Patil, S.R., Arnold, B.F., Salvatore, A.L., Briceno, B., Ganguly, S., Colford Jr., J.M., Gertler, P.J., 2014. The effect of India's total sanitation campaign on defecation behaviors and child health in rural Madhya Pradesh: a cluster randomized controlled trial. *PLoS Med.* (11), e1001709.
- Paul, J.R., Horstmann, D.M., Riordan, J.T., Opton, E.M., Niederman, J.C., et al., 1962. An oral poliovirus vaccine trial in Costa Rica. *Bull. World Health Organ.* 26, 311–329.
- Peletz, R., Cock-Esteb, A., Ysenburg, D., Haji, S., Khush, R., Dupas, P., 2016. The Supply and Demand for Improved Sanitation: Results from Randomized Pricing Experiments in Rural Tanzania (Working paper).
- Pitt, M.M., Rosenzweig, M.R., Hassan, N., 2012. Human capital investment and the gender division of labor in a brawn-based economy. *Am. Econ. Rev.* 102, 3531–3560.
- Pollitt, E., Hathirat, P., Kotchabhakadi, N., Missel, L., Valyasevi, A., 1989. Iron deficiency and education achievement in Thailand. *Am. J. Clin. Nutr.* 50 (3), 687–697.
- Pollitt, E., Gorman, K., Engle, P., Martorell, R., Rivera, J., 1993. Early Supplemental Feeding and Cognition. Monographs of the Society for Research in Child Development. Serial No. 235. University of Chicago Press, Chicago.
- Pop-Eleches, C., Thirumurthy, H., Habyarmina, J., Graff Zivin, J., Goldstein, M., DeWalque, D., MacKeen, L., Haberer, J., Sidle, J., Ngare, D., Bangsberg, D., 2011. Mobile phone technologies improve adherence to antiretroviral treatment in resource-limited settings: a randomized controlled trial of text message reminders. *AIDS* 25 (6), 825–834.
- Prina, S., Royer, H., 2014. The importance of parental knowledge and social norms: evidence from weight report cards in Mexico. *J. Health Econ.* 37, 232–247.
- Mathieu, S., Boutron, I., Moher, D., Altman, D.G., Ravaud, P., September 2, 2009. Comparison of registered and published primary outcomes in randomized controlled trials. *JAMA* 302 (9), 977–984.
- Robinson, J., Yeh, E., 2011. Transactional sex as a response to risk in western Kenya. *Am. Econ. J. Appl. Econ.* 3 (1), 35–64.
- Robinson, J., Yeh, E., 2012. Risk-coping through sexual networks: evidence from client transfers in Kenya. *J. Hum. Resour.* 47 (1), 107–145.
- Rosenbaum, P.R., 2007. Interference between units in randomized experiments. *J. Am. Stat. Assoc.* 102, 191–200. <http://dx.doi.org/10.1198/016214506000001112>.
- Rosenthal, R., 1979. The file drawer problem and tolerance for null results. *Psychol. Bull.* 86 (3), 638–641. <http://dx.doi.org/10.1037/0033-2909.86.3.638>.
- Rubin, D.B., 1990. Formal mode of statistical inference for causal effects. *J. Stat. Plan. Inference* 25, 279–292. [http://dx.doi.org/10.1016/0378-3758\(90\)90077-8](http://dx.doi.org/10.1016/0378-3758(90)90077-8).
- Schilbach, F., 2015. Alcohol and Self-Control: A Field Experiment in India. Harvard University (unpublished working paper).
- Seshadri, S., Gopaldas, T., 1989. Impact of iron supplementation on cognitive functions in preschool and school-aged children: the Indian experience. *Am. J. Clin. Nutr.* 50 (3), 675–686.
- Simonsohn, U., Nelson, L.D., Simmons, J.P., 2014. P-curve: a key to the file-drawer. *J. Exp. Psychol. General* 143 (2), 534–547.
- Singal, A.G., Higgins, P.D.R., Waljee, A.K., 2014. A primer on effectiveness and efficacy trials. *Clin. Transl. Gastroenterol.* 5, e45. <http://dx.doi.org/10.1038/ctg.2013.13>.
- Soemantri, A.G., Pollitt, E., Kim, I., 1989. Iron deficiency anemia and education achievement. *Am. J. Clin. Nutr.* 50 (3), 698–702.
- Soewondo, S., Husaini, M., Pollitt, E., 1989. Effects of iron deficiency on attention and learning processes of preschool children: Bandung, Indonesia. *Am. J. Clin. Nutr.* 50 (3), 667–674.
- Strauss, J., Thomas, D., 1995. Human resources: empirical modeling of household and family decisions. In: Behrman, J.R., Srinivasan, T.N. (Eds.), *Handbook of Development Economics*, vol. 3A. North Holland Press, Amsterdam.
- Sylvia, S., et al., 2015. Survey using incognito standardized patients shows poor quality care in China's rural clinics. *Health Policy Plan.* 30 (3), 322–333.

- Tarozzi, A., Mahajan, A., Blackburn, B., Kopf, D., Krishnan, L., Yoong, J., 2014. Micro-loans, insecticide-treated bednets and malaria: evidence from a randomized controlled trial in Orissa (India). *Am. Econ. Rev.* 104 (7), 1909–1941.
- Taylor-Robinson, D.C., Maayan, N., Soares-Weiser, K., Garner, P., 2012. Deworming drugs for treating soil-transmitted intestinal worms in children: effects on nutrition and school performance. *Cochrane Database Syst. Rev.* (7)
- The Voluntary HIV-1 Counseling and Testing Efficacy Study Group, 2000. Efficacy of voluntary HIV-1 counselling and testing in individuals and couples in Kenya, Tanzania, and Trinidad: a randomised trial. *Lancet* 356 (9224), 103–112. [http://dx.doi.org/10.1016/S0140-6736\(00\)02446-6](http://dx.doi.org/10.1016/S0140-6736(00)02446-6).
- Thirumurthy, H., Goldstein, M., Graff Zivin, J., 2008. The economic impact of AIDS treatment: labor supply in western Kenya. *J. Hum. Resour.* 43, 511–552.
- Thomas, D., et al., 2003. Iron Deficiency and the Well-Being of Older Adults: Early Results From a Randomized Nutrition Intervention. UCLA (unpublished manuscript).
- Thomas, D., et al., 2006. Causal Effect of Health on Labor Market Outcomes: Experimental Evidence. UCLA (unpublished manuscript).
- Thornton, R., 2008. The demand for, and impact of, learning HIV status. *Am. Econ. Rev.* 98 (5), 1829–1863.
- Vermeersch, C., Kremer, M., 2004. School Meals, Educational Achievement and School Competition: Evidence from a Randomized Evaluation. World Bank and Harvard University (unpublished working paper).
- World Bank, 2008. In: Amin, S., Das, J., Goldstein, M. (Eds.), *Are You Being Served?: New Tools for Measuring Service Delivery*, The World Bank, Washington, DC. <http://dx.doi.org/10.1596/978-0-8213-7185-5>.
- World Bank, 2015. *World Development Report 2015: Mind, Society and Behavior*. The World Bank, Washington, DC.
- World Development Report, 2004. Making Services Work for Poor People. World Bank. <https://openknowledge.worldbank.org/handle/10986/5986>.
- World Health Organization, 1993. *The Control of Schistosomiasis. Second Report of the WHO Expert Committee*. WHO, Geneva. Technical Report Series 830.
- World Health Organization, October 2015. Road Traffic Injuries. Fact sheet N°358. <http://www.who.int/mediacentre/factsheets/fs358/en/>.
- Ziegelhöfer, Z., 2012. Down with diarrhea: using fuzzy regression discontinuity design to link communal water supply with health. *Grad. Inst. Int. Dev. Stud. Work. Pap.* Available: <http://www.econstor.eu/handle/10419/77433>.
- Zivin, J.G., Thirumurthy, H., Goldstein, M., 2009. AIDS treatment and intrahousehold resource allocation: children's nutrition and schooling in Kenya. *J. Public Econ.* 93, 1008–1015.
- Zwane, A.P., Zinman, J., Van Dusen, E., Pariente, W., Null, C., Miguel, E., Kremer, M., Karlan, D., Hornbeck, R., Giné, X., Duflo, E., Devoto, F., Crepon, B., Banerjee, A., 2011. Being surveyed can change later behavior and related parameter estimates. *Proc. Natl. Acad. Sci.* 108.

## CHAPTER 2

# The Production of Human Capital in Developed Countries: Evidence From 196 Randomized Field Experiments<sup>a</sup>

R.G. Fryer, Jr. <sup>\*§,1</sup>

\*Harvard University, Cambridge, MA, United States

§NBER (National Bureau of Economic Research), Cambridge, MA, United States

<sup>1</sup>Corresponding author: E-mail: rffryer@fas.harvard.edu

## Contents

1. Introduction	96
2. A Method for Finding and Evaluating Field Experiments	105
3. Evidence From 196 Randomized Field Trials	110
3.1 Early childhood experiments	110
3.1.1 <i>Center-based experiments</i>	110
3.1.2 <i>Home-based experiments</i>	112
3.1.3 <i>Meta-analysis</i>	115
3.2 Home environment	116
3.2.1 <i>Parental involvement</i>	116
3.2.2 <i>Home educational resources</i>	120
3.2.3 <i>Poverty reduction experiments</i>	123
3.2.4 <i>Neighborhood quality</i>	127
3.2.5 <i>Meta-analysis</i>	129
3.3 Randomized field experiments in K-12 schools	129
3.3.1 <i>Student-based interventions</i>	130
3.3.2 <i>Teacher-based interventions</i>	140
3.3.3 <i>School Management</i>	157
3.3.4 <i>Market-based approaches</i>	161
4. Combining What Works: Evidence From a Randomized Field Experiment in Houston	169
4.1 Simulating the potential impact of implementing best practices in education on wage inequality	172
4.1.1 <i>Interpreting the literature through a simple life-cycle model</i>	174
4.1.2 <i>Simulating the Social Genome Model</i>	174

<sup>a</sup> I am grateful to Lawrence Katz and numerous colleagues whose ideas and collaborative work fill this chapter. William Murdock III provided a truly unprecedented amount of effort, attention to detail, and input into this project. Tanaya Devi and C. Adam Pfander also provided exceptional research assistance. Financial support from the Broad Foundation and the EdLabs Advisory Group is gratefully acknowledged. Correspondence can be addressed to the author by e-mail at rffryer@fas.harvard.edu. The usual caveat applies.

4.1.3 Simulating impacts on income	175
5. Conclusion	181
References	307

## Abstract

Randomized field experiments designed to better understand the production of human capital have increased exponentially over the past several decades. This chapter summarizes what we have learned about various partial derivatives of the human capital production function, what important partial derivatives are left to be estimated, and what—together—our collective efforts have taught us about how to produce human capital in developed countries. The chapter concludes with a back of the envelope simulation of how much of the racial wage gap in America might be accounted for if human capital policy focused on best practices gleaned from randomized field experiments.

## Keywords

Education Policy; Human Capital Production; Social Genome Model; Racial Achievement Gap; Meta-Analysis; Racial and Ethnic (Wage) Inequality; Randomized Field Experiments; Student Achievement.

## JEL Codes

C93; H52; H75; I21; I24; I25; I26; I28; I29; I38; J24; O15

*The True Method of Knowledge is Experiment*

*William Blake*

## 1. INTRODUCTION

Racial and ethnic inequality is a stubborn empirical reality across the developed world. Blacks in the United States earn 24% less, live five fewer years, and are six times more likely to be incarcerated on any given day ([Fryer, 2010](#)). Black men in the United Kingdom are three times more likely to be unemployed and as full-time workers, earn 20% less ([Hatton, 2011](#)). The Roma in Hungary are over two years less educated, have worse self-reported health, and earn 28% less ([Kántor, 2011](#)). Turkish immigrants in Germany are almost twice as likely to be unemployed and earn 38% less ([von Loeffelholz, 2011](#)). African immigrants in Spain are less educated than natives, have a 4.9 percentage point higher unemployment rate, and earn 35% less ([de la Rica, 2011](#)). The income difference between natives and second-generation immigrants in Sweden is 11% ([Nordin and Rooth, 2007](#)).

Gaining a better understanding of the underlying causes of such stark racial and ethnic inequality is of tremendous importance for public policy. Using data from the United States, [O'Neill \(1990\)](#) and [Neal and Johnson \(1996\)](#) demonstrate that blacks, Hispanics,

and whites are paid similar prices for similar premarket skill bundles—yet, there are large differences in skills. Similarly, [Nordin and Rooth \(2007\)](#) show that differences in income between natives and second-generation immigrants in Sweden depend strongly on a skill gap—when controlling for scores on the Swedish Military Enlistment Test, the income gap decreases by more than 70%.

An important question then is what obstacles preclude the acquisition of productive skills. Using 10 large datasets which together include students that range in age from 8 months to 17 years old, [Fryer and Levitt \(2013\)](#) show that the racial achievement gap is remarkably robust across time, samples, and assessments. The achievement gap does not exist in the first year of life, but black students in the United States fall behind by age 2 (in the raw data) and these racial differences in academic achievement after kindergarten cannot be explained by including standard controls. Similarly, controls cannot explain differences between children of natives and children of immigrants on international standardized tests such as the Programme for International Student Assessment (PISA), Trends in International Mathematics and Science Study (TIMSS), and Progress in International Reading Literacy Study in other developed countries such as France, Switzerland, Netherlands, and Sweden ([Parsons and Smeeding, 2008](#)).

If the deleterious effects of labor market discrimination are in decline and the importance of productive skills are on the rise, an important public policy question is how to increase human capital—particularly for those who, due to accident of birth, begin life disadvantaged. A fuller understanding would allow policymakers to use basic economic principles (e.g., equating marginal return to marginal costs) in their decision-making. Is it more cost effective to decrease class size or provide parents financial incentives to increase student achievement? Should school districts increase the management skills of principals or increase early childhood programs? What is a better use of resources—early childhood investments or providing high-dosage tutoring to adolescents?

In an effort to answer questions like these, education researchers have spent decades trying to infer causal relationships from nonexperimental data by examining large data sets and invoking various assumptions, many of which are not verifiable. Prior to the late 1970s, research on the relationship between class size and academic achievement was widely considered inconclusive ([Porwell, 1978; Glass and Smith, 1978](#)). In fact, some studies, including the famous Coleman Report, suggested there were greater gains in classrooms with *more* students ([Nelson, 1959; Coleman et al., 1966](#)). These studies did not adequately account for the fact that school districts commonly bundled better students and teachers in classrooms with more students. A well-designed randomized experiment would enable researchers to avoid such confounding factors and help settle the debate among nonexperimental estimates. Using the random assignment of students to small classes in Project STAR, [Krueger \(1999\)](#) showed that students

assigned to small classrooms indeed do score higher than students in regular-sized classrooms. The effect sizes for the K–3 students in Project STAR are in the range of 0.19–0.28 standard deviations and represent 64–82% of the white–black test score gap in the data.

Similarly, a large body of nonexperimental studies have found significant positive correlations between neighborhood socioeconomic status and students' academic achievement (Aaronson, 1998; Ainsworth, 2002; Chase-Lansdale and Gordon, 1996; Chase-Lansdale et al., 1997; Duncan et al., 1994; Halpern-Felsher et al., 1997; Kohen et al., 2002). However, randomized and quasiexperimental studies have failed to establish a causal link. Although Rosenbaum (1995) found that suburban students from Chicago's Gautreaux program outperformed urban students, Jacob (2004) found no effects on students' test scores from switching neighborhoods due to housing demolitions. Further, Oreopoulos (2003) found no evidence of long-term impacts of neighborhood quality on labor market outcomes in a quasiexperimental analysis. More importantly, in the short-run, the Moving to Opportunity randomized housing mobility experiment (Ludwig et al., 2012; Kling et al., 2007; Sanbonmatsu et al., 2011) produced no sustained improvements in academic achievement, educational attainment, risky behaviors, or labor market outcomes for either female or male children, including those who were below school age at the time of random assignment. Interestingly though, Chetty et al. (2016) show that the Moving to Opportunity experiment had large impacts on early-adulthood outcomes for children who were younger than 13 years old at randomization. In their mid-20s, these individuals have 31% higher income, have higher college attendance rates, are less likely to be single parents, and live in better neighborhoods relative to similar individuals in the control group. For children who were older than 13 years old at randomization, the experiment had no positive long-term impacts.

In the 1920s, William McCall, an education psychologist at Columbia University, was one of the first supporters of using randomization to investigate the validity of education programs. His 1923 book, "How to Experiment in Education," developed a method for gathering data by randomly determining treatment and control groups. His work provided the framework for the experimental designs we see in educational field experiments today. Many of the early influential education field experiments came decades after McCall's book with the wave of large-scale social experiments in the latter half of the 20th century.<sup>1</sup> In the 1960s we saw the Perry Preschool experiment and the income maintenance experiments, in the 1970s the Abecedarian project was initiated, and in the 1980s there was Project STAR, the Tennessee class size experiment. The data from these

<sup>1</sup> See Levitt and List (2009) for a brief history of field experiments.

randomized experiments alone were used for decades to investigate many interesting questions about how to best produce human capital.

The inherent power of randomized field experiments is in the ability to estimate partial derivatives of the educational production function. That is, holding other variables constant, one can vary experimentally the amount of time students spend in school or the salary of their teachers, or whether or not the students receive financial incentives. One's imagination is the only real bound.

To see the advantages of this approach, imagine the following simple production process.<sup>2</sup> Let  $Y_{ij}$  denote a measure of an academic achievement  $j$  for individual  $i$ , where  $j$  might represent state test scores or other norm-referenced tests such as the Peabody Picture Vocabulary Tests or the Woodcock–Johnson Tests of Achievement. For each  $j$ , assume a simple Education Production Function of the following form:

$$Y_i = f(E_i, S_i, H_i, M_i, P)$$

where,  $E_i$  denotes student  $i$ 's early childhood experience,  $S_i$  captures various school inputs,  $H_i$  represents household and neighborhood inputs,  $M_i$ , captures “social skills” such as grit, resilience, or what psychologists often refer to as “the Big 5.” Let  $P$  be a vector of relevant prices.

We assume that  $f$  is smooth and continuously differentiable in its arguments. Imagine that we want to understand the impact of important changes in home environment on student test scores, holding school quality, mindset, and early childhood experience fixed. This is equivalent to estimating  $\frac{\partial Y}{\partial H}$ . On the other hand, we may want to understand the impact of investments in school-based reform on human capital holding all else equal by estimating  $\frac{\partial Y}{\partial S}$ . Or, the impact of instilling more “grit” or a “growth mindset” into students, all else being equal. This is equivalent to  $\frac{\partial Y}{\partial M}$ .

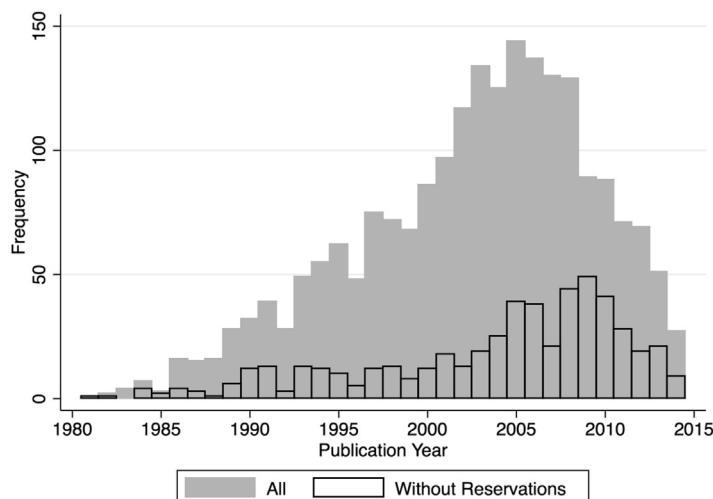
Perhaps recognizing the net benefits of randomized field experiments and because of a desire to avoid past miscues due to biased estimates, federal and local governments, early childhood centers, entrepreneurs, and school districts have become laboratories for randomized field experiments. Forty-five years after the famous Perry Preschool experiment, families in Chicago Heights were rewarded for teaching their own children a similar curriculum (Fryer et al., 2015a,b). Thirty years after the seminal class size experiment in public elementary schools of Tennessee, school districts in both America and Europe have implemented various tutoring experiments, management best practices,

<sup>2</sup> The model is meant to illuminate, clarify, and contrast estimates in the literature. It is not meant to be “realistic” or to be directly estimable. There is a rich literature designed to better understand and empirically estimate the education production function (Cunha and Heckman, 2007; Hanushek, 1979; Krueger, 1999; Todd and Wolpin, 2003).

and programs designed to increase the human capital of the adults in school buildings (e.g., Fryer, 2014b; Cook et al., 2014; Clark et al., 2013; Garet et al., 2008; Carlson et al., 2011; May et al., 2013; Blachman et al., 2004). Forty years after the income maintenance experiments, public policy across the developed world is being influenced by researchers investigating the impacts of welfare-to-work programs, earnings supplements, and parental involvement (e.g., Hamilton et al., 2001; Michalopoulos et al., 2002; Avvisati et al., 2014).

Indeed, randomized control trials in education have increased exponentially over the past 50 years. In 2000, 14% of reviewed education publications on What Works Clearinghouse met their standards without reservations, a distinction given only to well-designed studies that have comparison groups determined through a random process. By 2010, that number had tripled to over 46%. Fig. 1 provides a time series of studies in education. Throughout the 1980s, these randomized education studies were sparse. But in the 1990s, we start seeing a steady flow of approximately 10 publications a year that utilize a random design, and then this number increases all the way up to a high of 49 randomized experiments in 2009.

Given the remarkable increase in the use of randomized field trials over the past 50 years and the robust correlation between human capital and other economic outcomes such as income and employment, it is time to take stock and summarize what we have learned about various partial derivatives of the human capital production function, what important partial derivatives are left to be estimated, and



**Figure 1 Reviewed WWC studies.** This figure presents the number of reviewed studies in the What Works Clearinghouse (WWC) by publication year of the studies. The shaded histogram is the sample of all studies in WWC. The clear histogram is the sample of studies that met WWC's standards without reservation.

what—together—our collective effort over the past several decades has taught us about how to produce human capital in developed countries.<sup>3</sup>

This chapter attempts to do three things.

**First**, we conducted a relatively exhaustive search of all randomized field experiments in education. We define a field experiment as any intervention that uses a *verifiably* random procedure to assign participants to treatment and control groups in a nonlaboratory environment. This definition, while restrictive, is consistent with the definition of a field experiment described in [Harrison and List \(2004\)](#) and the US Department of Education's What Works Clearinghouse "without reservation" standard. Using this definition, we sourced almost 1000 field experiments to be included in our analysis. We further limited the sample of studies to be included to studies conducted in "highly developed" countries

<sup>3</sup> To be clear, randomized trials are not a panacea. There are important limitations to randomized controlled trials, which have been documented in [Deaton \(2010\)](#), [Mosteller and Boruch \(2002\)](#), [Worrall \(2007\)](#), and [Rothstein and von Wachter \(2017\)](#) in this volume. We describe a few here. First, many questions that are potentially interesting to economists may not be answerable with a randomized trial. For instance, how much of the variance in achievement is explained by genetic endowment? Given we are not likely to alter genetics by means of a field experiment, if one is wed to RCTs then this question is unanswerable. Second, as with all statistics—the evaluation of field experiments has implications for the mean of the population and may have little value in predicting individual behavior. With large enough RCTs, one can alleviate some of these concerns by estimating heterogeneous treatment effects. Third, and likely most constraining, are a host of important caveats which center on external validity. One cannot always generalize the results from a local RCT to other contexts. An obvious example of this is if an RCT finds a program has large impacts using a sample of poverty-stricken minority children, one cannot assume the program will have similar impacts on the universe of students in the United States. However, even if the RCT uses a representative sample of the target population, there are still concerns of external validity. For example, when implementing a large-scale policy, there could possibly be general equilibrium effects that a pilot RCT did not detect. Fourth, [Deaton \(2010\)](#) expresses many concerns about the analyses and implementations of RCTs—exploring heterogeneous treatment effects can be viewed as data mining and researchers should explore the implications of testing a large number of hypotheses in their studies; researchers rarely use appropriate standard errors when reporting results; exploring different combinations of baseline variables to include in regressions is another potential form of data mining; including baseline variables can lead to substantial biases in small samples; attrition from the study must be addressed; and it is not uncommon for RCTs to have implementation and operational issues that threaten the validity of the experiment. Fifth, spillover effects could lead one to misstate a program's overall effect. The example that [Rothstein and von Wachter \(2016\)](#) give is a labor market program that attempts to increase the search effort of individuals in treatment. This program may lower the chances of finding jobs for the control group and thus overstate the impact of the program's total effect. Sixth, RCTs evaluating programs are considered "black boxes" that do not reveal the true mechanisms of interest. Although one can use randomized admission lotteries to estimate the causal impact of preexisting charter schools, the causal relation between specific school inputs cannot be determined from such a study. Finally, [Deaton \(2010\)](#) and others argue that in an effort to overcome the above issues, RCTs can become prohibitively expensive. Still, with these important limitations in mind, the conventional wisdom is: if you *can* do a randomized field experiment, you should. Of the above seven issues which are commonly discussed with RCTs, five of them can be sidestepped by running more, larger, and better designed RCTs. Moreover, if one designs the RCT in a way that helps validate a model of selection for observational data, then the only limitation appears to be the budget of the researcher.

with standardized reading or math outcomes.<sup>4</sup> These restrictions eliminated almost three-quarters of the experiments, leaving a sample of 199.

We divide our sample of studies into three main categories of intervention—early childhood, school-based interventions, and home-based interventions—and provide a summary of the literature within each category.<sup>5</sup> Early childhood experiments investigate the impacts of preschool attendance, home-based initiatives that target prekindergarten children, and different preschool models on early achievement. Indeed, any experiment with outcomes measured before kids enter school is categorized as early childhood—*independent of the nature of the treatment*.

School-based experiments target K-12 curricula, teachers, management practices, students in classroom settings, principals, and other school resources. Any experiment where the dosage is applied in a school setting—such as offering families vouchers to attend private schools or after-school programs—we categorize as a school-based intervention. Even experiments in which K-12 resources are given at home—for instance tutors from the school tutor students in their living rooms—we code as a school-based experiment. Home-based experiments focus on parenting, income constraints, neighborhood environment, and a student's access to educational resources in their household. Similar to above, if an experiment takes place at home and focuses on these inputs, then it is considered a home-based experiment. For example, parenting classes that take place in a school auditorium are considered a home intervention.

While the above categories are mutually exclusive, collectively exhaustive, and internally consistent—which categories to sort experiments into is a bit arbitrary. For example, in [Sumi et al. \(2012\)](#), both teachers and parents received training on how to teach students replacement behaviors. This is potentially important because when we combine estimates within categories across the set of experiments using the DerSimonian–Laird metaanalysis coefficient (see [DerSimonian and Laird, 1986](#))—the labels on categories become a “lazy man’s” way of deciding what works and what does not. If the metaanalysis coefficient for early childhood studies is greater than the coefficient for home studies, this is evidence that early childhood studies have a higher impact on average. In an attempt to avoid interactions of the categories in our analysis, studies that have characteristics of more than one category are excluded from our analysis (but are still included in the tables).

<sup>4</sup> We consider countries as highly developed if they received a classification of “Very High Human Development” in [United Nations Development Programme \(2010\)](#). A country is classified as very high if they score in the top quartile on an index of human development that includes life expectancy, mean years of schooling, expected years of schooling, and gross national income per capita.

<sup>5</sup> We do not focus on mindset experiments ( $M_i$  in the production function above) due to very few of these experiments passing the inclusion restrictions of our meta-analysis discussed below.

With these caveats in mind, the results of this inquiry are interesting and, in some cases, quite surprising. There is substantial heterogeneity in treatment effects across and within various categories of field experiments. Experiments in early childhood and schools can be particularly effective at producing human capital. The random effects metacoeficients for early childhood experiments are  $0.111\sigma$  (0.031) for standardized math scores and  $0.165\sigma$  (0.032) for reading. The estimates for schools are  $0.052\sigma$  (0.008) and  $0.068\sigma$  (0.009) for math and reading scores, respectively. Within school-based field experiments, those that alter the management practices of schools, or implement “high-dosage” tutoring tend to demonstrate large effects. Having pooled impacts in the range of  $0.507$ – $0.655\sigma$ , the three most successful early childhood experiments were the famous Ypsilanti Perry Preschool Project (Weikart et al., 1970) and evaluations of the Breakthrough to Literacy and Ready, Set, Leap! curricula (Layzer et al., 2007). In schools, with evaluations producing pooled impacts ranging from 0.779 to  $1.582\sigma$ , the most successful programs appear to be Reading Recovery (Center et al., 1995; Schwartz, 2005) and Peer-Assisted Learning Strategies (Mathes and Babyak, 2001).

Interventions that attempt to lower poverty, change neighborhoods, or otherwise alter the home environment in which children are reared have produced surprisingly consistent and precisely estimated “zero” results. Avvisati et al. (2014) show that a comprehensive parent training program in France had large behavioral impacts that spilled over to students whose parents did not participate. However, the study found no impacts on academic outcomes. The famous negative income tax experiment—which provided low-income families with more money while incentivizing them to work less—had no impact on children’s test scores (Maynard and Murnane, 1979). As with Avvisati et al. (2014) and Maynard and Murnane (1979), the average home or neighborhood experiment that our search returned has math and reading impacts that are statistically indistinguishable from zero.

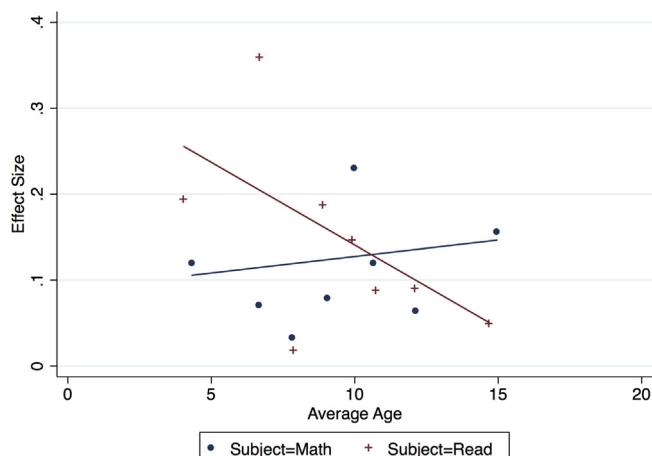
The literature—all 196 randomized field experiments discovered through our search process—is summarized in a large set of tables at the end of the chapter. This was the most laborious part of the process. For each study, we collected data on sample demographics, key aspects of the research design, and effect sizes. The typical study published in a top economics journal has this information readily available and collecting the data only took a few minutes. But, some studies published in older journals, less technical journals, or government reports required an exhaustive search of the publication to estimate effect sizes from the information given. The large collection of tables provides a bird’s eye view of the set of randomized field experiments that have been conducted and evaluated. We include a large set of studies that vary curriculum choices. These are included in the tables but not described in the text, as they do not align with traditional economic choice variables in a concise way and because of the potential effects of publication bias on these types of studies.<sup>6</sup>

<sup>6</sup> See Table A4.

**Second**, for every randomized field experiment found, we calculated the impact (in standard deviations) of the intervention on standardized math and reading outcomes and collected data on features of the experiment. These data include over 40 potential explanatory variables, including length of intervention, grade/age of subjects, location, if the sample was a majority English Language Learner (ELL), disadvantaged, black, Hispanic, or of low ability, and so on. This provides us with a novel dataset to investigate the correlation of important sample demographics and treatment effects of experiments designed to increase human capital.

An important pattern that arises in the data is the correlation between effectiveness of treatment and age of subjects at the time of intervention. It has also been observed that some interventions tend to be more effective at increasing math achievement relative to reading achievement (Fryer, 2014a; Abdulkadiroglu et al., 2011; Angrist et al., 2011; Dobbie and Fryer, 2011; Hoxby and Murarka, 2009; Gleason et al., 2010). There are many theories that may explain the disparity in treatment effects by subject area. A leading explanation posits that reading scores are influenced by the language spoken when students are outside of the classroom (Charity et al., 2004; Rickford, 1999). Charity et al. (2004) argue that if students speak nonstandard English at home and in their communities, increasing reading scores might be especially difficult. Research in developmental psychology has suggested a second possibility—that the critical period for language development occurs early in life, while the critical period for developing higher cognitive functions extends into adolescence (Hopkins and Bracht, 1975; Newport, 1990; Pinker, 1994; Nelson, 2000; Knudsen et al., 2006).

Our data suggest that the age theory has merit. In math, the treatment effect is not strongly related to the age of the student at the time of intervention (Fig. 2). The



**Figure 2 Correlation of effect size and average age of intervention.** This figure plots annual effect sizes versus average age of students in an intervention by subject. The sample includes all studies that passed our selection criteria for the meta-analysis. Each binned scatter plot was created by separating the data for the given subject into 8 equal-sized bins, computing the mean of the average age and effect sizes within each bin, then creating a scatter plot of these data points. The solid lines shows the best linear fit estimated on the underlying unbinned data estimated using a simple OLS regression.

correlation coefficient is 0.0679. In stark contrast, early in life, many reforms increase reading performance. Later in life, very few treatments have any effect on reading, save “high dosage” tutoring. Put precisely—there is a negative relationship between age and reading treatment effects. The correlation coefficient is  $-0.2069$ . To put this number in perspective, the average effect on reading of interventions targeting children with an average age less than 5 is  $0.177\sigma$ . The average effect of reading interventions targeting students with an average age greater than 14 is  $0.039\sigma$ .

**Third**, we conclude this chapter by simulating a life-cycle model that enables us to make an educated guess about how much of racial and ethnic wage inequality in America might be accounted for if we simply used the best practices gleaned from an exhaustive review of what works in the literature. Although a majority of the randomized field trials discussed in this chapter do not report impacts on adult outcomes, we are able to use correlations from the National Longitudinal Survey of Youth 1979 (NLSY79) and NLSY79 Children and Young Adults (CNLSY) datasets to simulate how shocks in a given life-stage will impact later outcomes. Specifically, we follow the methods described in [Winship and Owen \(2013\)](#) and construct a model similar to the Social Genome Model (SGM). Using this model, we estimate that if children were given a successful early childhood intervention and then received successful school-based interventions in mid-childhood and again in adolescence, one might dramatically reduce, and under some assumptions eliminate, wage inequality. Obviously, these types of educated guesses vis-à-vis simulations must be taken with a proverbial grain of salt.

The chapter proceeds as follows. [Section 2](#) describes our method for culling and standardizing field experiments from the education literature. [Section 3](#) describes evidence from randomized field experiments across the three categories: early childhood, home-based interventions, and school-based interventions. [Section 4](#) uses the estimates from the literature to simulate a life-cycle model and provide a sense of how much of racial wage inequality in America might be accounted for if government policy focused on the best practices gleaned from the literature. [Section 5](#) concludes. There are 100 pages worth of Appendix Tables that summarize the literature in a concise and consistent way.

## 2. A METHOD FOR FINDING AND EVALUATING FIELD EXPERIMENTS

In deciding which field experiments to include in our analysis, we first culled a reasonably exhaustive list of field experiments and then narrowed our focus to studies that satisfied certain criteria. We began by searching all “quick reviews” and “single study reviews” in the What Works Clearinghouse (WWC). WWC was created by the US Department of Education’s Institute of Education Sciences in 2002. Its goal is to provide reviews of education studies, policies, and interventions in order for researchers to determine “what works” in education. Currently, WWC has over 10,500 reviews available in an online searchable database. Eligible studies are reviewed by a team of WWC’s certified staff against WWC standards and assigned a rating. The highest rating of the Clearinghouse is reserved for studies that met standards without reservations. This implies that groups compared in the study were determined through a random process, there was low overall

attrition from the sample, the differential attrition across groups was low, and there were no confounding factors ([US Department of Education, 2015](#)).<sup>7</sup> Our search of WWC produced 115 randomized field experiments that met standards without reservations.

We augmented the WWC search by looking through recent education literature reviews (e.g., [Almond and Currie, 2010](#); [Fryer, 2010](#); [Heckman and Kautz, 2013](#); [Nye et al., 2006](#); [Yeager and Walton, 2011](#)) to ensure that all these potential studies had been included. In almost all cases, the randomized studies were already in the What Works Clearinghouse, but this process produced important additions.

Finally, we conducted relatively broad searches of known databases—such as ERIC, JSTOR, EconLit—that include education papers to augment our sample of studies. In each database, we searched for all phrases generated by concatenating one element from the set of strings (“early childhood,” “education,” “housing,” “neighborhood,” “parent,” “school,” “student,” “teacher”) with one element from (“experiment,” “random assignment,” “randomization”). For each database, we collected all hits that searching for these 24 unique phrases returned.<sup>8</sup> These searches provided us with over 10,000 citations to check. To conduct this laborious task, we had a team of five research assistants skim every article and select papers that explicitly mention a random process determining the experimental sample.

Using these approaches, we found 859 potential studies. [Table 1](#) describes how we narrowed our set of studies from 859 to 196. As discussed, we only included experiments that had samples determined by a verifiably random process that were precollege; that took place in a highly developed country (as determined by the Human Development Index constructed by the United Nations Development Programme); and that reported standardized reading or mathematics test scores as an outcome measure at posttest. The random process is important for causal inference—though the rise of strong quasiexperimental analyses makes this quite restrictive. Some experiments use non-norm referenced tests designed by the experimenter for the purpose of the experiment. These evaluations are not comparable across experiments and were omitted. In general, the restrictions are important for comparability and allow one to synthesize the estimates from the studies in the analysis below ([Table 2](#)).

Unfortunately, these restrictions lead to us not including some influential experimental studies. For example, our screening excluded the exploration of the impact of teacher value-added on students in [Chetty et al. \(2014\)](#) because their research design is nonexperimental. Housing demolitions in [Jacob \(2004\)](#) and the famous brown and

<sup>7</sup> That is, no factor other than the intervention itself is present that all treatment students in one group are exposed to and no students in the comparison group are exposed to. If a confounding factor is present, it would be impossible to distinguish between the effect of the intervention and the effect of the factor.

<sup>8</sup> JSTOR’s search algorithm occasionally returned thousands of results. Due to resource and time constraints, we decided to only collect the top 200 (as determined by “relevance”) results for each phrase in JSTOR.

**Table 1** Paper accounting

	<b>Number of papers (1)</b>
<b>Panel A: titles found</b>	
From broad search	≈ 8000
Selected for further review	859
Total included	196
<b>Panel B: reason for exclusion</b>	
College sample/outcomes	42
Design issues	96
Countries w/o very high HDI	57
Insufficient info	24
Paper not located	10
No standardized reading or math	356
Repeat paper	70
Sample issues	8

Notes: This table summarizes our search procedure for selecting papers for inclusion. Panel A displays the approximate number of titles our initial broad search returned, the number selected for further review, and the final sample of papers. Of the titles selected for further review, Panel B reports the number of papers that were excluded for the given reason. See Online Appendix A for details on each exclusion restriction.

blue eye experiments performed by Jane Elliot in her classrooms in the late 1960s also did not utilize verifiably random processes. A well-known incentive experiment in Israel ([Angrist and Lavy, 2009](#)) was excluded because the main outcome was receipt of matriculation certificates. Similarly, many important social-psychological, behavioral, and “mindset” experiments (e.g., [Mischel et al., 1972](#); [Cohen et al., 2006, 2009](#); [Wilson and Linville, 1982](#); [Aronson et al., 2002](#); [Miyake et al., 2010](#); [Duckworth et al., 2013](#)) were excluded because they did not report results for standardized math or reading outcomes or the sample was post high school.

For each experiment that passed our screening, we report estimates of the annual pooled effect sizes on reading and math outcomes, in standard deviations. If papers did not report results in this manner, we attempted to use the information given to calculate standardized effect sizes. For example, if impacts were presented as scale score points on a test, we would divide the coefficient by the standard deviation given in the summary statistics. The most common calculation we performed was using the average treatment and control posttest scores (or changes between pretest and posttest) as well as the corresponding standard deviations to calculate the standardized difference between the two groups.

Specifically, we used this information to calculate a statistic known as Hedge’s  $g$  and its corresponding standard error (see [Hedges, 1981](#) and [Lipsey and Wilson, 2000](#)). Since this

**Table 2** Meta-analysis

	Math			Reading		
	Unweighted average (1)	Fixed effects (2)	Random effects (3)	Unweighted average (4)	Fixed effects (5)	Random effects (6)
<b>Panel A: early childhood</b>						
All	0.120 (0.028)	0.111 (0.031) 20	0.111 (0.031)	0.202 (0.027)	0.106 (0.012) 44	0.189 (0.027)
<b>Panel B: home</b>						
All	0.039 (0.045)	-0.004 (0.008) 8	-0.004 (0.008)	0.078 (0.052)	0.010 (0.007) 22	0.010 (0.007)
Parental involvement	0.122 (0.115)	-0.001 (0.021) 3	-0.001 (0.021)	0.143 (0.103)	0.009 (0.021) 11	0.034 (0.050)
Educational resources	-0.060 (0.000)	-0.060 (0.050)	-0.060 (0.050)	0.072 (0.063)	0.015 (0.014) 7	0.015 (0.014)
Poverty reduction	0.008 (0.001)	0.008 (0.029) 2	0.008 (0.029)	0.022 (0.011)	0.016 (0.024) 4	0.016 (0.024)
<b>Panel C: school</b>						
All	0.135 (0.022)	0.035 (0.004) 72	0.053 (0.009)	0.203 (0.028)	0.023 (0.004)	0.069 (0.011)
Student incentives	0.039 (0.026)	0.016 (0.011) 5	0.024 (0.018)	0.097 (0.072)	0.016 (0.011) 8	0.021 (0.017)
High dosage tutoring	0.393 (0.095)	0.309 (0.106) 6	0.309 (0.106)	0.405 (0.047)	0.217 (0.030) 25	0.229 (0.033)
Low dosage tutoring	0.074 (0.045)	0.015 (0.013) 3	0.015 (0.013)	0.050 (0.045)	0.015 (0.015) 4	0.015 (0.015)
Teacher certification	0.031 (0.036)	0.028 (0.012) 5	0.030 (0.030)	0.000 (0.015)	0.007 (0.028) 3	0.007 (0.028)
Teacher incentives	0.052 (0.033)	0.002 (0.011) 7	0.022 (0.022)	-0.000 (0.021)	-0.006 (0.012) 4	-0.006 (0.012)

General PD	0.173 (0.075)	0.019 (0.024) 7	0.019 (0.024)	0.153 (0.060)	0.022 (0.023) 9	0.022 (0.023)
Managed PD	0.059 (0.009)	0.052 (0.016) 2	0.052 (0.016)	0.493 (0.187)	0.217 (0.029) 8	0.403 (0.120)
Data driven	0.107 (0.041)	0.043 (0.014) 4	0.057 (0.024)	0.071 (0.040)	0.009 (0.011) 4	0.030 (0.024)
Extended time	-0.033 (0.089)	0.019 (0.026) 4	-0.019 (0.068)	0.155 (0.136)	0.012 (0.029) 5	0.032 (0.048)
School choice/ vouchers	0.076 (0.035)	0.024 (0.018) 6	0.024 (0.018)	0.070 (0.040)	-0.010 (0.012) 7	0.023 (0.025)
Charters	0.121 (0.039)	0.088 (0.011) 9	0.110 (0.030)	0.072 (0.026)	0.038 (0.010) 9	0.048 (0.018)
No excuse charters	0.170 (0.048)	0.124 (0.022) 5	0.153 (0.042)	0.104 (0.040)	0.055 (0.018) 5	0.077 (0.031)

Notes: This table reports average effects for categories of papers discussed in the main text. Columns (1)–(3) report results for math estimates and columns (4)–(6) report results for reading estimates. Columns (1) and (4) report the unweighted average for the studies in a given category. Columns (2) and (5) report estimates from a fixed-effects meta-analysis. Columns (3) and (6) report estimates from a random-effects meta-analysis using the DerSimonian–Laird model (see [DerSimonian and Laird, 1986](#)). Panel A reports results for early childhood experiments. Panel B reports results for home experiments. Panel C reports results for school experiments. The first row of each panel reports the results for all studies included in the given panel. The sample includes all studies found that meet our inclusion restrictions and have annual impact estimates for the given subject. See the main text and Online Appendix A for details on our search procedure, inclusion restrictions, and the categories of papers. Standard errors are reported in parentheses. The number of observations is reported below the standard error.

measure is just the difference between the average test scores of treatment and control groups, point estimates obtained from this method are identical to intent-to-treat (ITT) estimates that do not include controls and use the same standard deviation to standardize the test scores. Note that since all studies included in this paper used a random procedure to assign treatment and control groups, point estimates from multivariable ITT regressions should not differ significantly from the raw differences. If possible, when necessary information for this statistic was missing we would make assumptions (e.g., equal number of students assigned to treatment and control or use the standard deviation from the national sample of the standardized test).<sup>9</sup> If there was not enough information presented in the paper for us to make credible assumptions, the study was excluded.

One common issue we encountered was the calculation of standard errors. Unfortunately, without having access to the microdata, it was not possible to calculate the appropriate standard errors for every effect size. In an attempt to not overstate the significance of an effect size, when calculating Hedges'  $g$ , we erred on the conservative side and used the number of units randomized to calculate the standard errors. For example, although [Slavin et al. \(1984\)](#) had a sample of 504 students, randomization was done at the school level ( $N = 6$ ).

### **3. EVIDENCE FROM 196 RANDOMIZED FIELD TRIALS**

#### **3.1 Early childhood experiments**

In the past 5 decades there have been many field experiments designed to increase achievement before kids enter school.<sup>10</sup> Appendix [Table A1](#) provides an overview of 44 randomized field experiments (from 24 papers), the ages they serve, and their treatment effects on standardized math and reading outcomes. Here, we partition the literature into interventions that are early childhood center-based and others that are more home-based.

##### **3.1.1 Center-based experiments**

Perhaps the most famous early intervention program for children involved 123 students in Ypsilanti, Michigan, who attended the Perry Preschool program in 1962 (58 were randomly assigned to treatment). The program consisted of a 2.5-hour daily preschool program and weekly home visits by teachers, and targeted children from disadvantaged socioeconomic backgrounds with IQ scores in the range of 70–85. An active learning curriculum—High/Scope—was used in the preschool program to support both the cognitive and noncognitive development of the children over the course of 2 years beginning when the children were 3 years old. [Schweinhart, Barnes, and Weikart](#)

<sup>9</sup> We documented all assumptions that were made for each study and these can be obtained from the author upon request.

<sup>10</sup> See [Carmiero and Heckman \(2003\)](#) or [Almond and Currie \(2010\)](#) for extensive reviews.

(1993) found that students in the Perry Preschool program had higher test scores between the ages of 5 and 27, 21% less grade retention or special services required, 21% higher graduation rates, and half the number of lifetime arrests in comparison to children in the control group. Considering the financial benefits that are associated with the positive outcomes of the Perry Preschool, Heckman et al. (2010) estimated that the rate of return on the program is between 7% and 10%, passing a traditional cost-benefit analysis.

Although an influential experiment, Heckman et al. (2009) argues that the randomization protocol for the Perry Preschool experiment was compromised. Post randomization, some children initially assigned to treatment whose parents were employed were swapped with control children whose parents were unemployed. The researchers' rationale for this swap was that employed mothers would find it difficult to participate in the home visits that treatment families received. Heckman et al. (2010) investigated the implications of these swaps and other potential issues with previously reported Perry results. Even after accounting for the compromised randomization (by correcting for the imbalance in preprogram variables and matching students), multiple-hypothesis testing, and small sample sizes of the original analysis, Heckman et al. (2010) still found statistically and economically significant impacts.

Another important center-based intervention, which was initiated 3 years after the Perry Preschool program is Head Start. Head Start is a preschool program funded by federal matching grants that is designed to serve 3- to 5-year-old children living at or below the federal poverty level.<sup>11</sup> The program varies across states in terms of the scope of services provided, with some centers providing full-day programs and others only half-day. In 2007, Head Start served over 900,000 children at an average annual cost of about \$7300 per child.

Evaluations of Head Start have often been difficult to perform due to the typical nonrandom nature of enrollment in the program.<sup>12</sup> Puma et al. (2010), in response to the 1998 reauthorization of Head Start, conducted an evaluation using randomized admission into Head Start.<sup>13</sup> The impact of being offered admission into Head Start

<sup>11</sup> Local Head Start agencies are able to extend coverage to those meeting other eligibility criteria, such as those with disabilities and those whose families report income between 100 and 130% of the federal poverty level.

<sup>12</sup> Currie and Thomas (1995) use a national sample of children and compare children who attended a Head Start program with siblings who did not attend Head Start, based on the assumption that examining effects within the family unit will reduce selection bias. They find that those children who attended Head Start scored higher on preschool vocabulary tests but that for black students, these gains were lost by age 10. Using the same analysis method with updated data, Garces et al. (2002) find several positive outcomes associated with Head Start attendance. They conclude that there is a positive effect from Head Start on the probability of attending college and—for whites—the probability of graduating from high school. For black children, Head Start led to a lower likelihood of being arrested or charged with a crime later in life.

<sup>13</sup> Students not chosen by lottery to participate in Head Start were not precluded from attending other high-quality early childhood centers. Roughly 90% of the treatment sample and 43% of the control sample attended center-based care.

for 3- and 4-year olds is 0.10–0.34 standard deviations in the areas of early language and literacy. For 3-year olds, there were also small positive effects in the social-emotional domain (0.13–0.18 standard deviations) and on overall health status (0.12 standard deviations). Yet, by the time the children who received Head Start services had completed first grade, almost all of the positive impact on initial school readiness had faded. The only remaining impacts in the cognitive domain are a 0.08 standard deviation increase in oral comprehension for 3-year-old participants and a 0.09 standard deviation increase in receptive vocabulary for the 4-year-old cohort ([Puma et al., 2010](#)).<sup>14</sup>

Other early childhood interventions—many based on the early success of Perry Preschool and Head Start—include the Abecedarian Project, the Early Training Project, the Milwaukee Project, and Tulsa’s universal prekindergarten program. The Abecedarian Project provided full-time, high-quality center-based child care services for four cohorts of children from low-income families from infancy through age 5 between 1971 and 1977. [Campbell and Ramey \(1994\)](#) find that at age 12, those children who were randomly assigned to the project scored 5 points higher on the Wechsler Intelligence Scale and 5–7 points higher on various subscales of the Woodcock–Johnson Psycho-Educational Battery achievement test.

### **3.1.2 Home-based experiments**

The most well known home-based field experiment in the early childhood years is the Nurse–Family Partnership. Through this program, low-income first-time mothers received home visits from a registered nurse beginning early in the pregnancy and continued until the child was two years old—a total of 50 visits over the first two years. The program aimed to encourage preventive health practices, reduce risky health behaviors, foster positive parenting practices, and improve the economic self-sufficiency of the family. In a study of the program in Denver in 1994–95, [Olds et al. \(2002\)](#) found that those children whose mothers had received home visits from nurses (but not those who received home visits from paraprofessionals) were less likely to display language delays and had superior mental development at age 2. In a long-term evaluation of the program, [Olds et al. \(1998\)](#) found that children born to women who received nurse home visits between 1978 and 1980 had fewer juvenile arrests, convictions, and violations of probation by age 15 than those whose mothers had not received treatment.

The Early Training Project provided children from low-income homes with summertime experiences and weekly home visits during the three summers before entering first grade in an attempt to improve the children’s school readiness. [Gray and Klaus \(1970\)](#) report that children who received these intervention services maintained higher Stanford–Binet IQ scores (2–5 points) at the end of 4th grade. The Infant Health

<sup>14</sup> The Early Head Start program, established in 1995 to provide community-based supplemental services to low-income families with infants and toddlers, had similar effects ([Administration for Children and Families, 2006](#)).

and Development Program specifically targeted families with low birth weight preterm infants and provided them with weekly home visits during the child's first year and biweekly visits through age 3, as well as enhanced early childhood educational care and bimonthly parent group meetings. [Brooks-Gunn et al. \(1992\)](#) report that this program had positive effects on language development at the end of first grade, with participant children scoring 0.09 standard deviations higher on receptive vocabulary and 0.08 standard deviations higher on oral comprehension. The Milwaukee Project targeted newborns born to women with IQs lower than 80; mothers received education, vocational rehabilitation, and child care training while their children received high-quality educational programming and three balanced meals daily at "infant stimulation centers" for 7 h a day, 5 days a week until the children were 6 years old. [Garber \(1988\)](#) finds that this program resulted in an increase of 23 points on the Stanford-Binet IQ test at age 6 for treatment children compared to control children.

Although the above parenting programs have shown promise, they are not widely accessible due to the time demands they place on parents and high implementation costs. [York and Loeb \(2014\)](#) investigate the impact of READY4K!, a low-cost text message program that targets parents of preschoolers. The program helps these parents support their children's literacy development by sending parents three text messages per week for an entire school year. These texts were designed to provide parents with information on the importance of their children developing particular skills, tips on how to support their children's development in a cost-effective manner, and encouragement. [York and Loeb \(2014\)](#) recruited parents from 31 preschool sites run by the San Francisco Unified School District's Early Education Department. Of the 874 eligible families, 440 enrolled and were randomly assigned to treatment group that participated in READY4K! or a control group.

At the end of the school year, [York and Loeb \(2014\)](#) collected survey responses from parents and teachers to investigate the intervention's impact on parental involvement. They found that treatment parents engaged in literacy activities at home with their child 0.22 to 0.34 standard deviations more than control parents and were 0.13–0.19 standard deviations more involved at preschool. To investigate the impact the intervention had on children's literacy development, [York and Loeb \(2014\)](#) collected scores from the Phonological Awareness Literacy Screening (PALS), a criterion-referenced test the school district administers to its early education students every spring.<sup>15</sup> They found that children in treatment families scored 0.344 standard deviations higher on the letter sounds subtest and 0.205 standard deviations higher on a measure of lower-case alphabet knowledge. However, there were no significant impacts on measures of name writing, uppercase letter knowledge, beginning word sounds, print and word awareness, rhyme awareness, and

<sup>15</sup> Note that since this study did not report results from a norm-referenced outcome, it was not included in our tables and analysis.

a summed score of all the PALS subtests. For the sample of students that progressed to higher level subtests of PALS, there were significant impacts on the uppercase letter subtest and the summed score. There was limited evidence that READY4K! had differential impacts across family characteristics.

[Fryer et al. \(2015a\)](#) conducted a parental incentive experiment in Chicago Heights—a prototypical low performing urban school district—by starting a parent academy that distributed nearly \$1 million to 257 families (these numbers include treatment and control). There were two treatment groups, which differed only in when families were rewarded, and a control group. Parents in the two treatment groups were paid for attendance at Parent Academy sessions—designed as information sessions to aid parents in educating their children—and for proof of their children’s homework completion and performance on benchmark assessments. The only difference between the two treatment groups is that parents in one group were paid in cash or via direct deposits (hereafter the “cash” condition) and parents in the second group received the majority of their incentive payments via deposits into a trust account which can only be accessed if and when the child enrolls in college (the “college” incentive condition). Eleven project managers and staff worked together to ensure that parents understood the particulars of the treatment; that the parent academy program was implemented with high fidelity; and that payments were distributed on time and accurately.

Across the entire sample, the impact on cognitive test scores of being offered a chance to participate in the parental incentive is  $0.119\sigma$  (with a standard error of 0.094). These estimates are nontrivial, but smaller in magnitude than some classroom-based interventions. For instance, the impact of Head Start on test scores is approximately  $0.145\sigma$ . The impact of the Perry Preschool intervention on achievement at 14 years old is  $0.203\sigma$ . Given the imprecision of the estimates, however, our results are statistically indistinguishable both from these programs and from zero. The impact of the “college” and “cash” incentive schemes are nearly identical.

[Fryer et al. \(2015a\)](#) report that the impact of being offered a chance to participate in our parental incentive scheme on noncognitive skills is large and statistically significant [ $0.203\sigma$  (0.083)]. These results are consistent with [Kautz et al. \(2014\)](#), who argue that parental investment is an important contributor to noncognitive development. Again, the “cash” and “college” schemes yield identical results.

They complement our main statistical analysis by estimating heterogeneous treatment effects across a variety of predetermined subsamples that we blocked on experimentally. Two stark patterns appear in the data. The first pattern is along racial lines: Hispanics (48% of the sample) and whites (8% of the sample) demonstrate large and significant increases in both cognitive and noncognitive domains. For instance, the impact of the parent academy for Hispanic children is  $0.367\sigma$  (0.133) on our cognitive score and  $0.428\sigma$  (0.122) on our noncognitive score. Among the small sample of whites, the impacts are  $0.932\sigma$  (0.353) on cognitive and  $0.821\sigma$  (0.181) on noncognitive. The identical estimates for blacks are actually negative but statistically insignificant on both cognitive and noncognitive dimensions:  $-0.234\sigma$  (0.134) and  $-0.059\sigma$  (0.129), respectively. Importantly,

*p*-values on the differences between races are statistically significant at conventional levels. We explore a range of possible hypotheses regarding the source of the racial differences (extent of engagement with the program, demographics, English proficiency, pretreatment scores), but none provide a convincing explanation of the complete effect.

The second pattern of heterogeneity in treatment that we observe in the data relates to pretreatment test scores. Students who enter our program below the median on noncognitive skills see no benefits from our intervention in either the cognitive or noncognitive domain. In stark contrast, students who enter our parent academy above the median in noncognitive skills experience treatment effects of roughly 0.3 standard deviations on both cognitive and noncognitive dimensions. If we segment children by both cognitive and noncognitive pretreatment scores, the greatest gains are made on both the cognitive and noncognitive dimensions by students who start the program above the median on noncognitive skills and below the median on cognitive skills.

### 3.1.3 Meta-analysis

Early childhood interventions have amassed considerable popular and political support. Yet, like other initiatives to improve human capital, they are not a panacea. For example, St. Pierre et al. (1997) find no positive effects in a national evaluation of the Comprehensive Child Development Program (CCDP). The CCDP delivers early and comprehensive services to low-income families with the aim of enhancing the development of the children in these families and helping the parents achieve economic self-sufficiency. The CCDP model revolves around the ideas that one should intervene as early as possible in children's lives, involve the entire family in an intervention, deliver comprehensive services to address the needs of young children, enhance parents' ability to contribute to their child's development, help parents achieve economic and social self-sufficiency, and ensure that families have access to all of these resources until their children enter elementary school. In their evaluation, St. Pierre et al. (1997) found no significant differences between families that were randomly assigned to a CCDP treatment group or a control group. CCDP had no impact on measures of mothers' economic self-sufficiency or their parenting skills, and CCDP had no effects on the cognitive or social emotional development of the children included in the study.

Still, early childhood investments are considered to be one of the least risky ways to increase academic achievement (Heckman, 2008). Combining the 44 randomized studies in early childhood over the past 50 years, the random effects coefficients are  $0.111\sigma$  ( $0.031$ ) for math interventions and  $0.189\sigma$  ( $0.027$ ) for reading. Of the 64 treatment effects recorded in these randomized studies, 21 were statistically positive; none was statistically negative and 43 were statistically indistinguishable from zero.<sup>16</sup>

<sup>16</sup> We consider an effect size statistically positive or negative if it is statistically significant at the 10% level.

## 3.2 Home environment

There is an ongoing debate as to whether efficient production of human capital should focus on improving the environment in which a child lives or the environment in which they learn. Proponents of the school-centered approach refer to anecdotes of excellence in particular schools or examples of other countries where poor children in superior schools outperform average Americans (Chenoweth, 2007). Advocates of the community-focused approach argue that teachers and school administrators are dealing with issues that originate outside the classroom, citing research that shows racial and socioeconomic achievement gaps are formed before children ever enter school (Fryer and Levitt, 2004, 2006), that mother's IQ is highly correlated with child achievement (Fryer and Levitt, 2013; Yeates et al., 1983) and that one-third to one-half of the racial achievement gap can be explained by family-environment indicators (Phillips et al., 1998; Fryer and Levitt, 2004). In this scenario, combating poverty and having more constructive out-of-school time may lead to better and more-focused instruction in school. Indeed, Coleman et al. (1966), in their famous report on the equality of educational opportunity, argue that schools alone cannot treat the problem of chronic under-achievement in urban schools.

In this subsection, we describe several attempts to provide households with more resources and to combat poverty, in an effort to increase student achievement. We organize this strand of literature in rough approximation to the “intensity” of treatment received—which ranges from providing parents with information to poverty reduction through welfare-to-work programs and tax reform to moving families to better neighborhoods. The literature is summarized in Appendix Table A2.

### 3.2.1 Parental involvement

Parents matter. Using data from a national, cross-sectional study of children aged 8–12, Davis-Kean (2005) found significant correlations between parents' characteristics and parenting practices and students' math and reading achievement. Specifically, Davis-Kean (2005) found that strong correlations with students' achievement existed for parents' education levels, income, parental expectations, number of books owned, and many parental behaviors such as being warm and affectionate, responding positively, and giving praise. Jeynes (2005) conducts a meta-analysis of 41 studies that investigate the impact of parental involvement on the academic achievement of elementary students. He found that increases in parental involvement have an effect size on elementary students' academic outcomes of about 0.7 standard deviations. Jeynes (2007) conducts a similar meta-analysis using 52 studies that focus on secondary school students and finds the effect size of parent involvement to be about 0.5 standard deviations.

Although these results are interesting, they are not causal estimates of the impact of parental involvement on students' outcomes. Levels of parental involvement are most

likely correlated with many observable and unobservable characteristics of the parents and it is exceedingly difficult to rid these estimates of thorny issues of selection. Moreover, even if these estimates were causal, it is not obvious that it is possible for interventions to change parents' involvement to achieve these positive impacts on child outcomes.

In what follows, we summarize the literature on experiments to increase parental involvement using information treatments and incentive treatments.

### 3.2.1.1 Information

To better understand the impact of parental attitudes and school involvement on student achievement, [Avvisati et al. \(2014\)](#) conducted an experimental study on middle school students and parents in the educational district of Creteil, an eastern suburb of Paris, France. Classrooms randomly selected from 34 middle schools were offered a parental education program that taught parents how they can assist in their child's educational process.

This paper was motivated by a strong perception that disadvantaged parents have inadequate knowledge and confidence to be effective advocates for their children. The experiment sought to test if this could be improved by a simple intervention. The experimental program consisted of three after-school meetings with parents, conducted by the school head. The first two sessions focused on how parents can help their children's education by participating at home and at school. The final session, which took place after the end-of-term report card, focused on how parents can adapt to their children's first term results. Of the 352 state-run middle schools in the Creteil district, 34 schools volunteered to participate in the program. Around two-thirds of schools in the study were "priority education," a label indicating a historically disadvantaged area.

Parents of sixth graders in the participating middle schools were asked, over a 6-week period, if they would like to sign up for the informational meetings. After the sign-up period closed, the list of registered families constituted the "volunteer families," creating two populations within each class in each school. There were no strong observable pre-treatment differences between volunteer and nonvolunteer families. After registration closed, randomization began at the class-level of each school (meaning that roughly half of all classes were treated within each school). The randomization process defined four basic groups of families within each school: volunteers in treatment classes, nonvolunteers in treatment classes, volunteers in control classes, and nonvolunteers in control classes.

The study was interested in addressing three outcomes: (1) parental involvement attitudes and behavior; (2) children's behavior as reflected by truancy, disciplinary record, and work effort; and (3) children's academic results. To measure parental involvement attitudes and behavior, all families received a questionnaire on school-based involvement, home-based involvement, and parents' perception of the school. Student

outcomes were reported by teachers and academic reports. Main subject teachers were also given a questionnaire regarding both parental attitudes and child's behavior/school performance.

The evidence found that the program was successful in significantly improving volunteer parent attitudes. Based on parents' and teachers' questionnaires, parental involvement by volunteer parents in treatment classes increased. Children of volunteer parents in treatment classes saw a vast improvement in school attitudes and discipline compared to control classes: truancy was lower by 1.1 half-days, treatment students were 4.6 percentage points less likely to be punished for disciplinary reasons (6.4% versus 11.0%), more likely to earn top marks for conduct and, according to teacher questionnaire answers, were more likely to be agreeable in class and work diligently. In addition to having a direct impact on the students whose parents volunteered to participate, there were also spillover effects on students in treatment classrooms whose parents did not participate. Treatment had a statistically significant impact on nonvolunteer students' absenteeism, probability of disciplinary sanctions, and marks for conduct. For students of volunteer parents, treatment increased average grades across all subjects by 0.08 standard deviations and increased academic performance as measured by the teacher survey. However, the intervention had no impact on grades for students whose parents did not volunteer and the intervention had no impacts on standardized test scores for any students. The findings overall suggested that parental involvement can be a significant input in student achievement—mostly through an impact on behavioral outcomes.

Evidence from [Avvisati et al. \(2014\)](#) and other studies suggest that it may be difficult to increase students' academic outcomes using parental interventions. Other parental experiments that focus on improving students' academic outcomes through parental tutoring also tend to have insignificant impacts on academic standardized measures ([Warren, 2009](#); [Powell-Smith et al., 2000](#); [Fantuzzo et al., 1995](#); [Hirst, 1972](#); [Ryan, 1964](#)).

On average, parental information experiments increased student achievement by  $-0.001\sigma$  (0.021) on math scores and  $0.034\sigma$  (0.050) on reading scores. Note that our search did not return any parental incentive experiments that focused solely on parents of K-12 students. Therefore, the estimates from our meta-analysis for parental involvement and parental information are identical.

### 3.2.1.2 Incentives

The most well-known and well-analyzed incentive program for parents is PROGRESA. PROGRESA was an experiment conducted in Mexico in 1998, which provided cash incentives linked to health, nutrition, and education. The largest component of PROGRESA was linked to school attendance and enrollment. The program provided cash payments to mothers in targeted households to keep their children in school ([Skoufias, 2005](#)). Programs based on the PROGRESA model have been replicated in New York City, Nicaragua, and Columbia.

Beginning in 1997, the Mexican government identified 506 rural communities on the basis of a “marginality index” gleaned from census data. Socioeconomic data were collected from households within these communities to target households living in extreme poverty. In 1998, about two-thirds of the identified localities were randomly selected to receive financial incentives under PROGRESA; the remaining localities served as controls. As a part of the program, households could receive up to \$62.50 per month if children attended school regularly. The amount of incentive was higher for older children who had to attend 85% of all school days. The average amount of incentives received by any treatment household in the first 2 years of treatment was \$34.80, which was 21% of an average household’s income. Besides school attendance, PROGRESA also emphasized actual student achievement by making a child ineligible for the program if she or he failed a grade more than once ([Skoufias, 2005](#); [Slavin, 2010](#)).

[Schultz \(2000\)](#) reports that PROGRESA had a positive impact on school enrollment for both boys and girls in primary and secondary schools. For primary school children, PROGRESA increased school enrollment for boys by 1.1 percentage points and 1.5 percentage points for girls from a baseline level of approximately 90%. For secondary school students, enrollment increased by 7.2–9.3 percentage points for boys and 3.5–5.8 percentage points for girls, from a baseline level of approximately 70%. The author also reports that PROGRESA had an accumulated effect of 0.66 years additional schooling for a student from the average poor household. Taking the baseline level of schooling at face value, PROGRESA’s 0.66 years accumulated effect translates into a 10% increase in schooling attainment.

[Behrman et al. \(2001\)](#) also analyze the data and report that PROGRESA children entered school at an earlier age, had less grade repetition and better grade progression. Treatment children also had lower dropout rates and once dropped out, they had a higher chance of re-entry into high school.

Opportunity NYC—based on PROGRESA—was an experimental conditional cash transfer program that was conducted in New York City. The program had three components: the Family Rewards component that gave incentives to parents to fulfill responsibilities towards their children; the Work Rewards component that gave incentives for families to work; and the Spark component that gave incentives to students to increase achievement scores in classes. The program began in August 2007 and ended in August 2010 (see [Morais de Sá e Silva, 2008](#)).

[Riccio et al. \(2013\)](#) analyze data from the Family Rewards component of the program during the first 2 years of treatment. Their analysis is based on 4800 families with 11,000 children out of which half were assigned to treatment and the other half to control. Opportunity NYC spent \$8700 per family in treatment over 3 years. The experiment had an insignificant impact on every school outcome measured ([Riccio et al., 2013](#)).

### **3.2.2 Home educational resources**

Education, like other industries, has evolved over the past few decades—due, in part, to technological change. With the introduction of computers, the internet, mobile Wi-Fi, and smart phones, teaching strategies have changed to utilize these technologies in the classroom. However, many children still lack access to these resources in their homes. One could imagine that the returns to household computers are quite high. Students can use them as a tool to efficiently complete assignments, learn new information, study, and use for other educational purposes. Despite these potential returns, it is also possible that households face constraints (e.g., credit or information) that prevent them from investing in household technology. This is supported by the fact that ownership of household computers and access to household internet is correlated with income ([National Telecommunication and Information Administration, 2011](#)). Studies examining the impact of home computers on poor families using observational or quasiexperimental data have generated mixed results. Some studies find large positive effects ([Attewell and Battle, 1999](#); [Fiorini, 2010](#); [Schmitt and Wadsworth, 2006](#); [Fairlie, 2005](#); [Fairlie et al., 2010](#); [Malamud and Pop-Eleches, 2011](#)) and some find evidence of small or even negative impacts ([Fuchs and Woessmann, 2004](#); [Vigdor and Ladd, 2010](#); [Malamud and Pop-Eleches, 2011](#)). [Fairlie and Robinson \(2013\)](#) present causal estimates from the first ever randomized control experiment that investigates the impact of home computers.

In their experiment, Fairlie and Robinson investigate the educational impacts of randomly giving home computers to 1123 students in grades 6 through 10 in California over the 2008–2009 and 2009–2010 school years. No students who participated in the study had home computers at baseline. Half of these students were randomly selected to receive free computers without any training or technological assistance. Fairlie and Robinson collected administrative data on student academic outcomes and demographics pretreatment and at the end of the school year (posttreatment). In addition, they conducted baseline and posttreatment surveys that included questions about computer usage, knowledge, homework time, and other important outcomes. Using this data, Fairlie and Robinson found that the experiment had large first-stage impacts. They found that treatment students were 55% points more likely to have a computer at follow-up, 25% points more likely to have Internet service, they reported using a computer 2.5 h more per week than control students' average of 4.2 h, and almost all of this additional usage came from a computer at home. However, not all of the computer usage was for educational purposes. Relative to control students, treatment students used computers 0.80 h more per week for schoolwork (control mean (CM) was 1.89 h), 0.42 h more for e-mail (CM = 0.25 h), 0.80 h more for games (CM = 0.84 h), and 0.57 h more for social networking (CM = 0.57 h).

Despite these large first-stage impacts, Fairlie and Robinson find minimal evidence for impacts on educational outcomes. ITT estimates for the impact of home computers

on grades in math, English/reading, social studies, and science classes are all close to zero and precisely estimated. With standard errors of approximately 0.04, they can rule out effect sizes on the scale of one-fourth of the difference between a “B+” or “B” with 95% confidence. Using quantile regressions, they show that these null effects exist across the entire posttreatment achievement distribution. Similarly, they find no evidence of impact on students’ test scores or proficiency statuses from the California Standardized Testing and Reporting (STAR) program, total credits taken in the third quarter of the school year, total credits in the fourth quarter, unexcused absences, number of tardies, and if a student was still enrolled at the end of the school year. These zero effects are consistent with survey results that show treatment students did not change intermediate inputs and outcomes such as school effort, computer knowledge, and usage of important educational software.

There are other randomized field experiments that investigate the impact of providing additional resources to families—such as giving students books to read during the summer. Numerous studies suggest that summer vacation is a critical time for forming and widening achievement gaps in reading, particularly for the income-achievement gap.<sup>17</sup> Kim (2005) conducted an experimental study to examine the causal effects of a voluntary summer reading intervention on the reading skills of fourth-grade students in the Lake County Public School District, a large multiethnic school district located in a mid-Atlantic state. The district contains more than 100 elementary schools and is therefore organized into small subdistricts, each with its own superintendent. To be included in the sample, the subdistrict needed to contain high-poverty schools that administered Title I school-wide programs and contain multiracial schools in which reading scores for black and Latino students contributed to the federal adequate yearly progress rating. The final sample included four Title I schools and the six non-Title I schools with the largest percentage of minority students.

This paper was motivated by inefficiencies in current voluntary reading policies and the little evidence in support of these programs. Additionally, finding a cost-effective reading intervention was important for policymakers and practitioners given the goals of federal education policy and mandates under the No Child Left Behind Act.<sup>18</sup> The intervention addressed three main factors—access to books, students’ reading levels, and students’ reading preferences—that are likely to shape opportunities to read in the summer and affect reading outcomes. To increase access to books, each student in the treatment group received eight free books to read during the summer. Students’ reading levels were based on performance on the reading section of the Iowa Test of Basic Skills

<sup>17</sup> See Heyns (1978), Cooper et al. (1996), Alexander et al. (2001), Broh (2004), Heyns (1987), Klibanoff and Haggart (1981), Murnane (1975), and Phillips et al. (1998). Fryer and Levitt (2004) is a notable example of a nationally representative sample that does not find “summer setback.”

<sup>18</sup> Note that the No Child Left Behind Act was superseded by the Every Student Succeeds Act in December 2015.

and preferences were obtained through a survey distributed before the summer. A text-leveling system, the Lexile Framework, was used to provide books that were within each student's independent reading level using information about each student's reading level and reading preferences. With each book, students also received a postcard that asked students to check comprehension strategies used while reading the book and to obtain a signature from a parent or family member after reading a portion of the book aloud to the adult. Parents were instructed to mail each postcard back to the schools, regardless of whether their student completed the book or not.

A total of 552 students received consent to participate in the study and took pretests in June 2005. These students were randomly assigned to treatment and control groups (282 treatment and 270 control) within their English Language Arts (ELA) classroom, and the author reports no statistically significant differences between the two groups at the beginning of the experiment on numerous demographic and achievement characteristics. Because of attrition, the final sample included 486 students (252 treatment and 234 control) at the beginning of the Fall in 2005. The intervention attempts to improve reading skills by increasing children's access to books, matching books to children's reading levels and preferences, and encouraging children to read orally with a parent/family member to practice.

To investigate if the intervention increased children's access to books at home and literacy-related activities during summer vacation, the author used a two-way ANOVA on both self-reported measures of book ownership and on literacy habits gathered from a survey conducted at the end of the summer. The results suggest that the intervention did not increase children's access to books nor the amount of silent reading. However, children in the treatment group reported significantly more oral-reading at home with family members than the control group children. For fall reading outcomes, ITT regressions showed no significant differences between the treatment and control groups on a grade level measure of oral-reading fluency. However, treatment had a  $0.08\sigma$  (0.04) impact on students' standardized reading test scores and there were differential effects by race. Treatment increased test scores by  $0.22\sigma$  (0.09) for black students,  $0.14\sigma$  (0.08) for Latino students, and  $0.17\sigma$  (0.11) for Asian students. Further, the magnitude of the treatment effect was largest among lower performing students, and there were no significant interactions between the treatment and measures of reading ability or ownership of books.

Similarly, [Allington et al. \(2010\)](#) conducted a randomized trial in 17 high-poverty schools in Florida where treatment students selected 12 books from a book fair to receive for summer reading. [Allington et al. \(2010\)](#) found a  $0.046\sigma$  (0.033) annual impact over the 3 years of the experiment. The metacoeficients for home educational resource experiment were  $-0.060\sigma$  (0.050) for math scores and  $0.015\sigma$  (0.014) for reading scores.

### **3.2.3 Poverty reduction experiments**

One of the most often articulated explanations for the racial and ethnic achievement gaps that exist across developed countries is poverty. For families with higher income, it is easier to provide their children with resources and raise them in environments that are conducive for learning. Poverty places constraints on key factors of achievement such as health care, nutrition, child care, in-home educational resources, safe neighborhoods, good schools, and college education (Brooks-Gunn and Duncan, 1997; Evans, 2004; Magnuson and Duncan, 2002; McLoyd, 1998). In America, 42% of black children and 37% of Hispanic children experience poverty while only 10% of white children are exposed to these hardships (Duncan and Magnuson, 2005). Studies suggest that this racial income gap is an important source of variation that can account for large proportions of raw racial achievement gaps (Duncan and Magnuson, 2005; Fryer and Levitt, 2004; Phillips et al., 1998; Brooks-Gunn et al., 2003).

This subsection discusses the impact on student achievement of experiments that attempted to reduce poverty through tax reform and work programs.

#### **3.2.3.1 Tax reform**

Maynard and Murnane (1979) discuss two mechanisms by which welfare reform could affect children's educational achievement by altering home environment: product inputs and time inputs. Product inputs are things such as food, health care, and books. Examples of time inputs are time parents spend talking to, playing with, and reading to their children. They assume that product inputs are positively related to family income and that time inputs are positively related to time not working. Maynard and Murnane investigate the educational impacts of a program that affects both of these mechanisms in unison by increasing families' income and incentivizing them not to work.

In the early 1970s, the Gary Income Maintenance Experiment was conducted by Indiana University under contracts with the US Department of Health, Education, and Welfare and the Indiana State Department of Public Welfare. Families who voluntarily enrolled and had at least one child under the age of 18 were randomly assigned to negative income tax conditions or control. Of the 1799 eligible families, 57% were assigned to one of four negative income tax plans for three years. These tax plans were a combination of two tax rates (40 or 60%) and two guaranteed income levels (about three fourths of the poverty level or equal to the poverty level). The lower guarantee level was about \$1000 a year more than the support level of the Indiana Aid to Families with Dependent Children program.<sup>19</sup> The tax rate is the amount by which the negative income tax payment is reduced for each dollar of income that a family earns.

<sup>19</sup> In 1972, the official poverty level for a four-person nonfarm family was \$4275.

The sample for the Gary Income Maintenance Experiment was not nationally representative. All children were black and three-fifths of them lived in female-headed households. In addition, the average family had a much lower income compared to the national average (the average annual income of families in the Gary experiment was only \$5200 and the national average at that time was \$9433) and over 40% of the Gary families were living below the poverty line.

Maynard and Murnane investigate the impact of a Gary family's assignment to any one of the treatment arms on educational outcomes of students in grades 4–10 at the end of the experiment (three years after randomization). They found that treatment increased students' standardized reading test scores by 0.23 standard deviations on the Iowa Test of Basic Skills in grades 4–6 but had no significant impact for students in grades 7–10. They found no evidence that treatment had an effect on grade point average of the younger students, but found that it significantly decreased grade point average for the older students. They also found no evidence that treatment had an impact on the number of days absent for either group of students.

To better understand these results, Maynard and Murnane also investigate the mechanisms by which the experiment might have affected school performance. They found that the Gary experiment had a significant first stage impact on total family income, but caused minimal change in the number of hours worked. Treatment families on average had their incomes increased by \$2000 per year (approximately a 50% increase). For married mothers, there was no change in hours worked per week. For female family heads, there was a decrease of about 2 h per week. Additionally, experimental families that lived in public housing before randomization were more likely to move to private dwellings than control families that lived in public housing prior to randomization. However, there was no statistical difference in mobility in the pooled sample.

### 3.2.3.2 Work programs

[Michalopoulos et al. \(2002\)](#) evaluated another poverty reduction program, called the Self-Sufficiency Project (SSP) that attempted to make work more appealing than welfare to long-time welfare recipients in the Canadian provinces of British Columbia and New Brunswick by providing them with wage subsidies. New Brunswick is located in eastern Canada and is bordered by the U.S. state of Maine on its western boundary. New Brunswick has a population of 750,000, a majority of its inhabitants speak English as their first language, and has a per capita GDP of 42,600 Canadian dollars. British Columbia is located in western Canada and is bordered by the U.S. states of Alaska, Washington, Idaho, and Montana. British Columbia has a population of 4,400,000, an official language of English, and a per capita GDP of 47,500 Canadian dollars.<sup>20</sup>

<sup>20</sup> Statistics come from the 2011 Canadian census (Statistics Canada 2013).

The study randomly assigned 6000 single parents from British Columbia and New Brunswick, who had been on income assistance for at least one year, to a treatment and control group. Treatment parents were eligible to participate in SSP and control parents were not. Parents enrolled in SSP received a monthly earnings supplement conditional on starting a full-time job and leaving income assistance. The earnings supplement was in addition to earnings from employment for up three years, as long as the parent continued to be employed full-time and remained off of income assistance. After random assignment, treatment parents had one year to find full-time employment (at least 30 h per week) and leave income assistance to enroll in SSP. After enrollment, the supplement participants received was half of the difference between their earnings and an earnings benchmark (the benchmark varied by location and year, but was \$30,000 in New Brunswick and \$37,000 in British Columbia for the first year of the experiment). This supplement was not affected by unearned income, earnings of other family members, and number of children. This supplement would essentially double the wage of many low-wage workers.

[Michalopoulos et al. \(2002\)](#) found significant first-stage impacts. Thirty-six percent of single parents that were offered participation found full-time employment and took-up the supplement during the year long eligibility window. Of those that participated in SSP, the average parent received the supplement for 22 months over the 3 years of the program and received more than \$18,000 in supplements over that at that time. SSP increased treatment parents' probability of employment throughout the duration of the program and reduced income assistance payments received by these families. As a result, treatment parents earned nearly \$3400 more than control members. Total income (supplements, earnings, and income assistance) increased by \$6300 for the average treatment family. These impacts reduced the proportion of treatment parents below Canada's low income cut-offs by 10 percentage points. Although these large impacts were observed during the program, these impacts did not persist after the completion of SSP. By 6 years after random assignment (2 years after all treatment parents would have stopped receiving supplements), treatment and control parents were equally likely to be employed and had similar average earnings.

[Michalopoulos et al. \(2002\)](#) also investigated the impact of SSP on the outcomes of the parents' children. They found differential treatment effects by the age of the child at the beginning of treatment. For children who were 1 or 2 years old at the time of random assignment, SSP had no effects on their performance on a standardized test of vocabulary skills (Peabody Picture Vocabulary Test) and achievement as reported by parents. For children who were 3 or 4, SSP increased students' scores on a math skills test and parental-reported achievement. Treatment children who were 13, 14, or 15 at the time of random assignment reported doing worse in school and committing more minor acts of delinquency during the program, but these effects faded away after parents were no longer eligible for the supplement. Finally, for older adolescents, SSP had no impacts

on educational, crime, or work related outcomes, but these students were significantly more likely to have babies. Other than the effects stated above for the young adolescents, there was no evidence of SSP having any impacts on health, behavior, and the emotional well-being of students in the study.

In a large analysis of welfare-to-work programs in the US, [Hamilton et al. \(2001\)](#) conduct a national evaluation of the long-term effects of 11 welfare-to-work programs on the recipients and their children. The evaluation investigates the effectiveness of two different types of preemployment strategies, Labor Force Attachment (LFA) and Human Capital Development (HCD). LFA welfare-to-work programs typically consist of short-term job search and encourage welfare participants to find employment quickly. HCD programs emphasize investment in longer-term skills and typically encourage participants to enroll in training or basic education programs. [Hamilton et al. \(2001\)](#) use data on over 40,000 single parents (mostly female) and their children who were randomly assigned to these programs in sites across the nation to investigate the impact of LFA and HCD programs.

Over the course of the 5-year follow-up period, a majority of control group members worked at some point. For example, 88% of the control parents from the Grand Rapids site were employed at some point. In Oklahoma City, 79% worked at some point during that time and 66% worked in Riverside. Although there was a high percentage of control parents that ever worked, treatment parents still worked during more calendar quarters on average than control parents in 9 of 11 programs. Similarly, in 9 of 11 programs, treatment parents on average had higher total earnings. Typically, [Hamilton et al. \(2001\)](#) found that employment-focused programs produced employment and earnings effects almost immediately while education focused programs did not have effects until a year or more after randomization. However, when directly comparing the LFA and HCD programs in the sites where they were run side by side, employment and earnings levels over the 5 years were very similar.

By the end of the follow-up period, almost all control families were off of welfare and the average control group member remained on assistance for only 2–3 years. However, both treatment types still reduced months on welfare relative to the control averages and there is some evidence that LFA treatment members left welfare assistance at a faster pace than HCD participants. These reductions in welfare usage appear to directly offset the increase in salary. Despite increasing earnings, treatment largely had no impact on total combined income (earnings, welfare and Food Stamp payments, and Earned Income Tax Credits).

[Hamilton et al. \(2001\)](#) also investigate if the welfare-to-work programs had effects on family circumstances and children's well-being. They found that there was no evidence of impacts on health care coverage, marriage rates, and few impacts on household composition and living arrangements. However, adults assigned to a welfare-to-work program were less likely to report recent physical abuse at the end of the experiment.

To investigate impacts on children, the researchers conducted a Child Outcomes Study in six of the programs (three different sites that each offered LFA and HCD

programs). These studies included almost 50 measures of children's academic functioning, health, social skills, and behavior for children who were preschool age at randomization. The authors report that 15% of these tests produced statistically significant differences, but the sign and magnitudes were rarely consistent across sites. For example, the estimates from the Atlanta LFA and HCD programs suggested favorable impacts on social skills and behavior for young children, but the Grand Rapids programs revealed negative effects. For older children, the programs led to few significant results. However, whenever results for these students were significant, they tended to be unfavorable. For example, an HCD program at one site increased the likelihood of dropping out, increased percentage of adolescents who had a physical, emotional, or mental condition that impeded their mother's ability to go to work, and increased teenage pregnancies among families with lower levels of education. Note again that no effects varied consistently by program approach or site for adolescents. Summarizing the literature on poverty reduction attempts to increase student achievement, the metacoeficients for this strand of literature are  $0.008\sigma$  (0.029) and  $0.016\sigma$  (0.024) on math and reading respectively. And, perhaps more telling, there is not one experiment that generates statistically significant positive effects on standardized test scores.<sup>21</sup>

### **3.2.4 Neighborhood quality**

A more nuanced version of the “poverty is first-order” argument is that the mechanism by which disadvantage affects achievement is not directly through income—hence, addressing the income problem has no real impact—but through what sociologists refer to as “a culture of poverty.” This theory argues that the poor are not simply lacking resources, but are also immersed in a culture that develops mechanisms or has social institutions that perpetuate poverty (Moynihan, 1969; Harrington, 1982). Taking the culture of poverty paradigm at face value, the randomized field experiment that one would ideally conduct would be to move families from high-poverty to low-poverty neighborhoods—particularly when children are young. This is precisely what the Moving to Opportunity (MTO) randomized housing mobility experiment did—one of the most pathbreaking experiments of our generation.

From 1994 to 1998, MTO enrolled 4604 poor families with children residing in public housing in high-poverty neighborhoods of Baltimore, Boston, Chicago, Los Angeles, and New York City. Families were randomly assigned to three groups: (1) the experimental voucher group, which received a restricted housing voucher that could be used to pay for private rental housing initially restricted to be in a low-poverty area (a census tract with under a 10% poverty rate in 1990) and some housing-mobility counseling; (2) the Section 8 only voucher group, which received

<sup>21</sup> However, note that some of these studies found impacts for subsamples of the participants or on noncognitive outcomes.

regular Section 8 housing vouchers with no MTO relocation constraint; and (3) a control group, which received no assistance through MTO. Across the MTO treatment sites, 61% of household heads were non-Hispanic blacks, 31% were Hispanic, and nearly all households were female-headed at baseline. About half of the experimental voucher group and 63% of the Section 8-only voucher group were able to obtain leases and move with an MTO voucher (the compliance rate). The MTO families were tracked for 15 years using administrative data as well as major interim (4–7 years after random assignment) and long-term (10–15 years after random assignment) follow-up surveys and analyses (Kling et al., 2007; Sanbonmatsu et al., 2011). MTO generated large and persistent improvements in residential neighborhoods for the treatment groups (especially the experimental voucher group) relative to the control group but only modest changes in school quality. The average MTO family lived at baseline in a neighborhood with a 53% poverty rate. MTO led to a 9 percentage point decline in the duration-weighted average tract poverty rate over the 10- to 15-year follow-up period for the experimental voucher group relative to the control group.

In stark contrast, MTO only modestly improved school quality for the MTO treatment groups. From the time of random assignment until the long-term follow-up, the experimental voucher group children attended schools that outperformed their control group peers by only 3 percentile points on state exams, and the Section 8 only voucher group children attended schools that performed just 1 percentile point higher. MTO treatment group students also typically remained in schools where the majority of the students were low-income and minority. MTO reduced the share of students eligible for free or reduced-price lunch by 4 percentage points for the experimental voucher group. Although it is difficult to compare the size of the neighborhood quality change to that of the school quality change, MTO appears to have elicited a larger improvement on neighborhood quality. The MTO treatment groups experienced more than twice as large a reduction in the share of poor residential peers as compared to poor school peers and more than three times as large an improvement in percentile rank in the national census-tract poverty distribution for their neighborhoods than in the state test score distribution for their schools. Many of the MTO movers remained in the same school districts and very similar schools. MTO also had no significant impact on adult economic self-sufficiency or family income at the interim or long-run follow-ups. Thus an analysis of the impacts of MTO treatments on child outcomes comes close to getting at the pure effects of changes in home and neighborhood conditions for disadvantaged kids (with little change in schools or family economic resources):  $\frac{\partial Y}{\partial H}$  in our framework.

The MTO voucher treatments did not detectably impact parent's economic outcomes, but they did significantly and persistently improve key aspects of mother's (adult female's) mental and physical health including substantial reductions in psychological distress, extreme obesity, and diabetes (Ludwig et al., 2011; Sanbonmatsu et al., 2011). MTO movers also experienced significant increases in adult subjective well-being with

larger gains for adults from sites where treatment induced larger reductions in neighborhood poverty (Ludwig et al., 2012). For female youth, MTO treatments similarly led to persistent and significant improvements in mental health (including substantial reductions in psychological distress) and marginally significant improvements in physical health, but there were no long-term detectable health impacts for male youth (Kling et al., 2007; Sanbonmatsu et al., 2011). Analyses 4 to 7 and 10 to 15 years after randomization found that MTO produced no sustained improvements in academic achievement, educational attainment, risky behaviors, or labor market outcomes for either female or male children, including those who were below school age at the time of random assignment. Interestingly though, using administrative data from tax returns through 2012, Chetty et al. (2016) show that the Moving to Opportunity experiment has had large impacts on early-adulthood outcomes for children who were younger than 13 years old at randomization. In their mid-20s, these individuals have 31% higher incomes; have higher college attendance rates; are less likely to be single parents; and live in better neighborhoods relative to similar individuals in the control group. For children who were older than 13 years of age at randomization, MTO had no positive long-term impacts.

The MTO findings imply that large improvements in neighborhood conditions for poor families (at least in the range feasible with Section 8 housing vouchers) alone do not produce noticeable gains in children's short-term socioeconomic and educational outcomes but can have substantial impacts on important long-term outcomes for children who were exposed to these environment changes before the age of 13. The lack of school-quality changes, induced by treatment, is suggestive of a key role for schools in children's short-term educational outcomes and risky behaviors.

### 3.2.5 Meta-analysis

Combining all the randomized studies for home environment, the random effects coefficients are  $-0.004\sigma$  (0.008) for math interventions and  $0.010\sigma$  (0.007) for reading. Astonishingly, the only study that had a statistically positive pooled impact was an unpublished dissertation. These results show that interventions that directly impact parents and households have struggled to have immediate effects on students' achievement outcomes.

## 3.3 Randomized field experiments in K-12 schools

Thus far, the literature suggests that early childhood experiments yield strong effects, but policies designed to reduce poverty, increase work opportunities, or increase neighborhood quality do little to affect the production of human capital of schoolchildren. In this section, we explore 105 randomized field experiments conducted in K-12 schools. The literature is summarized in Appendix Table A3.

We categorize experiments into four buckets: student-based interventions, teacher-based interventions, management reforms, and “market-based” reforms.

### **3.3.1 Student-based interventions**

#### **3.3.1.1 Financial incentives**

Perhaps the most natural way to increase human capital production—at least to an economist—is to change the incentives of schoolchildren to exert effort. Of course, rational agents—even little ones—internalize the returns to education that accrue in the labor market. Yet, if agents discount the future or are otherwise “boundedly rational,” individual effort may be below the optimum. Financial incentives offer a chance to bridge the gap and thereby increase effort.

There is a nascent but growing body of scholarship on the role of incentives in primary, secondary, and postsecondary education around the globe ([Angrist et al., 2002](#); [Angrist and Lavy, 2009](#); [Kremer et al., 2009](#); [Behrman et al., 2005](#); [Angrist et al., 2006](#); [Angrist et al., 2009](#); [Fryer, 2011](#); [Fryer and Holden, 2013](#); [Barrera-Osorio et al., 2011](#); [Bettinger, 2012](#); [Hahn et al., 1994](#); [Jackson, 2010](#)). We describe a subset of the literature below.

**3.3.1.1.1 Incentives in primary schools** Psychologists argue that children understand the concept of money as a medium of exchange at a very young age ([Marshall and MacGruder, 1960](#)), but the use of financial incentives to motivate primary school students is exceedingly rare.<sup>22</sup> [Bettinger \(2012\)](#), who evaluates a pay-for-performance program for students in grades 3 through 6 in Coshocton, Ohio, is one notable exception. Coshocton is 94% white and 55% free/reduced-price lunch. Students in grades 3 through 6 took achievement tests in five different subjects: math, reading, writing, science, and social studies. [Bettinger \(2012\)](#) reports a  $0.13\sigma$  increase in math scores and no significant effects on reading, social science, or science. Pooling subjects produces an insignificant effect.

[Fryer \(2011\)](#) and [Fryer and Holden \(2013\)](#) also describe student financial incentive experiments that target primary students that were conducted during the 2007–2008 and 2010–2011 school year in Dallas and Houston respectively. In Dallas, [Fryer \(2011\)](#) paid 2nd graders \$2 per book to read and pass a short computer-based comprehension quiz on the book in Accelerated Reader (AR), a software program that has quizzes for 80,000 trade books, all major reading textbooks, and leading children’s magazines. Students were allowed to select and read books of their choice at the appropriate reading level and at their leisure, not as a classroom assignment. The books came from the existing stock available at their school (in the library or in the classroom). To reduce the possibility of cheating, quizzes were taken in the library on a computer and students were

<sup>22</sup> The use of nonfinancial incentives—gold stars, aromatic stickers, certificates, and so on—are a more common form of incentive for young children. Perhaps the most famous national incentive program is the Pizza Hut Book It! Program which provides one-topping personal pan pizzas for student readers. This program has been in existence for 25 years, but never credibly evaluated.

only allowed one chance to take a quiz. Data on the number of books read for students in control schools in Dallas was not available because control schools did not have consistent access to AR. In total, the experiment distributed \$42,800 (21,400 quizzes passed) to 1777 children across the 21 treatment schools.

Paying students to read books yielded a treatment effect of  $0.012\sigma$  (0.069) in reading and  $0.079\sigma$  (0.086) in math. The key result from this analysis emerges when one partitions students in Dallas into two groups based on whether they took the exam administered to students in bilingual classes (Logramos) or the exam administered to students in regular classes (Iowa Test of Basic Skills). Splitting the data in this way reveals that there is a  $0.173\sigma$  (0.069) increase in reading achievement among English speaking students and a  $0.118\sigma$  (0.104) decrease in reading achievement among students in bilingual classes. When we aggregate the results in our main analysis this heterogeneity cancels itself out. Similarly, the treatment effect for students who are not English Language Learners is  $0.221\sigma$  (0.068) and  $-0.164\sigma$  (0.095) for students who are English Language Learners.

[Fryer and Holden \(2013\)](#) conducted a randomized field experiment in 50 traditionally low-performing public schools in Houston, Texas—providing financial incentives to fifth grade students, their parents, and their teachers in 25 treatment schools. Students received \$2 per math objective mastered in Accelerated Math (AM), a software program that provides practice and assessment of leveled math objectives to complement a primary math curriculum. Students practice AM objectives independently or with assistance on paper worksheets that are scored electronically and verify mastery by taking a computerized test independently at school. Parents also received \$2 for each objective their child mastered and \$20 per parent—teacher conference attended to discuss their student’s math performance. Teachers earned \$6 for each parent—teacher conference held and up to \$10,100 in performance bonuses for student achievement on standardized tests. In total, the experiment distributed \$51,358 to 46 teachers, \$430,986 to 1821 parents, and \$393,038 to 1734 students across the 25 treatment schools.

The experimental results raise a number of questions. On outcomes for which direct incentives were provided, there were very large and statistically significant treatment effects. Students in treatment schools mastered  $1.087\sigma$  (0.031) more math objectives than control students. On average, treatment parents attended almost twice as many parent—teacher conferences as control group parents. And, perhaps most important, these behaviors translated into a  $0.081\sigma$  (0.025) increase in math achievement on Texas’s statewide student assessment. The impact of our incentive scheme on reading achievement (which was not incentivized) is  $-0.077\sigma$  (0.027), however, offsetting the positive math effect. These results are consistent with the classic multitasking and job design work of [Holmstrom and Milgrom \(1991\)](#).

Interestingly, there is significant heterogeneity in treatment effects as a function of pretreatment test scores. Higher-achieving students (measured from pretreatment test

scores) master  $1.66\sigma$  more objectives, have parents who attend two more parent–teacher conferences, have  $0.228\sigma$  higher standardized math test scores and equal reading scores relative to high-achieving students in control schools. Conversely, lower-achieving students master  $0.686\sigma$  more objectives, have parents who attend 1.5 more parent–teacher conferences, have equal math test scores and  $0.165\sigma$  *lower* reading scores. Put differently, higher-achieving students put in significant effort and were rewarded for that effort in math without a deleterious impact in reading. Lower-achieving students also increased effort on the incentivized task, but did not increase their math scores and their reading scores decreased significantly. These data suggest that the classic “substitution effect” may depend on baseline ability.

Two years after removing the incentives, the treatment effect for high-achieving students is large and statistically significant in math [ $0.271\sigma$  ( $0.110$ )] and is small and statistically insignificant in reading. In stark contrast, low-achieving students have no treatment effect in math but a large, negative, and statistically significant treatment effect on reading [ $-0.219\sigma$  ( $0.084$ )]. These data suggests that there may be long-run impacts of multitasking through learning, dynamic complementarities, or both.

**3.3.1.1.2 Incentives in secondary schools** Fryer (2011) and Fryer (2010) describe the results of a series of randomized field experiments on financial incentives and secondary student achievement. In NYC, seventh grade students were paid for performance on a series of 10 interim assessments administered by the NYC Department of Education to all students. In Chicago, ninth graders were paid every 5 weeks for grades in their core courses. In Washington, DC, sixth, seventh, and eighth grade students were paid for their performance on a metric that included attendance, behavior, and three inputs to the production function chosen by each school individually.

The results reported in Fryer (2011, 2010) are surprising. The impact of financial incentives on state test scores is statistically zero in each city. In NYC, paying students for performance on standardized tests yielded treatment effects of  $0.004\sigma$  ( $0.017$ ) in reading and  $-0.031\sigma$  ( $0.037$ ) in mathematics in seventh grade and similar results for fourth graders. In Chicago, rewarding ninth graders for their grades had no effect on achievement test scores in math or reading. In Washington, DC, where students were paid for various inputs to the educational production function, we observed an impact of  $0.152\sigma$  ( $0.092$ ) in reading and  $0.114\sigma$  ( $0.106$ ) in mathematics.

Overall, these estimates suggest that incentives are not a panacea—but we cannot rule out small to modest effects (e.g.,  $0.10\sigma$ ) which, given the relatively low cost of providing financial incentives to students, have a positive return on investment.

Perhaps even more surprisingly, financial incentives had little or no effect on the outcomes for which students received direct incentives, self-reported effort, or intrinsic motivation. In NYC, the effect of student incentives on the interim assessments is, if anything, negative. In Chicago, where we rewarded students for grades in five core subjects,

the grade point average in these subjects increased  $0.093\sigma$  (0.057) and treatment students earned 1.979 (1.169) more credits (half a class) than control students. Both of these impacts are marginally significant. Incentives in Washington DC had no significant impacts on attendance rates, report card grades, or behavioral incidents.

Treatment effects on an index of “effort,” which aggregates responses to survey questions such as how often students complete their homework or ask their teachers for help, are small and statistically insignificant across all cities, though there may have been substitution between tasks. Finally, using the Intrinsic Motivation Inventory developed in [Ryan \(1982\)](#), [Fryer \(2011, 2010\)](#) and [Fryer and Holden \(2013\)](#) find little evidence that incentives decrease intrinsic motivation.

Taken together, the randomized field experiments involving financial incentives for students have generated a rich set of facts. Paying 2nd grade students to read books significantly increases reading achievement for students who take the English tests or those who are not English Language Learners, and is detrimental to non-English speakers. Paying fifth graders for completing math homework significantly increases their math achievement and significantly decreases their reading achievement. All other incentive schemes had, at best, small to modest effects—none of which were statistically significant.

### 3.3.1.2 Nonfinancial incentives and returns to schooling

[Fryer \(2013a\)](#) describes a large and innovative randomized field experiment which grew out of a partnership between three large organizations: Tracphone—the largest prepaid mobile phone provider in the United States, Droga5—an internationally recognized advertising firm, and the Oklahoma City Public Schools. The experiment, entitled “The Million,” was designed to provide accurate information to students about the importance of education on future outcomes such as unemployment, incarceration, and wages and to provide incentives to read books through free cell phones and minutes to talk and text.

Students in three treatment groups were given cellular phones free of charge, which came preloaded with 300 credits that could be used to make calls or send text messages. Students in the main treatment arm received 200 credits per month to use as they wanted and received one text message per day on the link between human capital and future outcomes delivered at approximately 6:00 p.m. A second treatment arm provided the same text messages as well as nonfinancial incentives—credits to talk and text were earned by reading books outside of school. A third treatment arm allowed students to earn credits by reading books and included no information. There was also a pure control group that received neither free cellular phones, information, nor incentives.

On direct outcomes for students in the informational treatments, [Fryer \(2013a\)](#) reports students’ ability to answer specific questions about the information provided in the text messages. Treatment effects were uniformly positive. Pooling across both informational treatments, treatment students were 4.9 (2.7) percentage points more likely to

correctly identify the wage gap between college graduates and college dropouts, 17.9 (3.8) percentage points more likely to correctly identify the relationship between schooling and incarceration, and 17.8 (3.8) percentage points more likely to answer both questions correctly. As a robustness test, we included a “placebo” question on the unemployment rate of college graduates, about which students never received information. The difference in the probability of answering this question correctly between informational treatments and the control group was trivial and statistically insignificant. Moreover, 54% of control students believe that incarceration rates for high school graduates and dropouts are “no[t] differen[t]” or “really close,” suggesting that students in Oklahoma Public Schools do not have accurate knowledge of the returns to schooling.

Results are mixed for indirect outcomes such as self-reported effort, state test scores, and attendance. Across the treatment arms, ITT estimates of the effect of treatment on self-reported effort are positive and statistically significant for both incentives and information arms. For instance, students in the information treatment were 15.1 (3.7) percentage points more likely to report feeling more focused or excited about doing well in school and 7.0 (3.7) percentage points more likely to believe that students were working harder in school.

In stark contrast, on all administrative outcomes—math or ELA test scores, student attendance, or behavioral incidence—there was no evidence that any treatment had a statistically significant impact, though due to imprecise estimates one cannot rule out small-to-moderate effects which might have a positive return on investment.

Another potentially powerful incentive is offering students a chance to earn college credit or college degrees while still in high school. The idea is that offering college credit will increase student incentives to exert effort and increase access to college for some students. Over 240 schools nationwide, called Early Colleges, have already adopted this model. Early Colleges combine a rigorous high school curriculum along with the potential to earn 2 years of college credit or a 2-year degree during high school. Most Early Colleges target underserved students and team up with colleges to offer this opportunity at no or low cost to the students. Berger et al. (2013) utilize the random lottery admission process of some Early Colleges to investigate the causal impact of Early Colleges on students’ outcomes.

In their study, Berger et al. (2013) used administrative and survey data from 10 Early Colleges that conducted random admission lotteries for the 2005–06, 2006–07, or 2007–08 school years. Comparing lottery winners to lottery losers, they were able to estimate causal impacts on high school completion, college enrollment, college degrees earned, standardized test scores, and high school and college experiences. High school outcome and student demographic data were obtained directly from the administrative records of the schools involved in the study; for college outcomes, students were matched to records in National Student Clearinghouse; data on high school and college experiences as well as college credits obtained while in high school came from a student survey that the researchers administered to students in 8 of the Early Colleges. The final sample

included 2458 students for the administrative outcomes and 1294 students for the survey outcomes.

Using this data, Berger et al. (2013) found that students offered admission to Early Colleges were significantly more likely to graduate high school and ever attend college than students who lost the lottery. Eighty-six percent of Early College students graduated high school compared to 81% of lottery losers, and 80% of Early College students enrolled in college whereas only 71% of comparison students did. Note that these numbers only reflect enrollment observed during the study period, 2005–2011, and that the gap in enrollment rates between lottery winners and lottery losers was decreasing as time went on. For example, for cohorts with 6 years of data available, the gap 4 years after ninth grade was 39.2 percentage points and this gap had decreased to 9.8 percentage points 6 years out. Further, when restricting the sample to only students that enrolled in college after high school graduation, lottery students were only 5.7 percentage points more likely to attend a 4-year college and they find no significant differences for any college or 2-year college enrollment. Similarly, during the study period, Early College students were 20 percentage points more likely to obtain a college degree (control mean was 2%). These degrees were typically associate's degrees and approximately 20% of Early College students earned a degree before the end of high school.

Berger et al. (2013) found no impact on GPA and math standardized test scores. However, they found that Early College students scored 0.14 standard deviations higher than lottery losers on standardized ELA tests. The survey results showed that Early College students were 45.1 percentage points more likely to earn college credit in high school than comparison students and that comparison students were 33.7 percentage points more likely to take at least one advanced placement exam in high school. In addition, Early College students reported engaging in rigorous learning activities in school more frequently, being exposed to higher expectations of college attendance from teachers, principals, and their peers, and reported receiving more help for completing college applications and financial aid forms.

The findings overall suggest that Early Colleges can successfully impact students' college enrollment and attainment during the 4 years that the students are enrolled in an Early College—but that these impacts might not spillover to the years following high school graduation.

The metaanalysis coefficients for student incentives experiments are  $0.024\sigma$  (0.018) for math achievement and  $0.021\sigma$  (0.017) for reading.

### 3.3.1.3 Tutoring

Throughout recorded history, the children of the elites were taught in a manner that would now be referred to as tutoring. In ancient Greece, children from wealthy families received their primary education individually or in small groups from masters or tutors (Dunstan, 2010). This practice continued for children of the rich and nobles throughout

the Middle Ages ([Nelson-Royes, 2015](#)). As late as the 17th century, schooling was thought to be a social, not academic, activity with primary human capital produced in small groups at home. Yet, the term, “tutoring,” has in more recent history become synonymous with remediation and fallen out of favor.

There is substantial heterogeneity in how schools implement various programs that fall under the general umbrella of “tutoring.” Some schools place students in one-on-one settings with a trained tutor, other schools place eight students with a volunteer. Some students receive tutoring 30 min per week; others are provided 5 h of intense instruction in the same time period. This heterogeneity leads, naturally, to large differences in treatment effects. [Fryer and Dobbie \(2013\)](#), define “high-dosage” tutoring as being tutored in groups of 6 or fewer for 4 or more days per week. Moreover, they demonstrate that tutoring itself is not correlated with charter school effectiveness. However, schools who implement “high-dosage” tutoring demonstrate marked treatment effects.

Following [Fryer and Dobbie \(2013\)](#), we divide the randomized field experiments in tutoring into these two groups: low-dosage and high-dosage tutoring. They are discussed in turn. In this exposition, high-dosage tutoring is defined as being tutored in groups of 6 or fewer for more than 3 days per week or being tutored at a rate that would equate to 50 h or moreover a 36-week period.<sup>23</sup>

**3.3.1.3.1 High-dosage tutoring** [Blachman et al. \(2004\)](#) report results from a study of a high-dosage tutoring program that targeted struggling second and third grade readers. Their study specifically focused on these young readers in an attempt to increase the growth trajectories of these students and possibly combat the negative adolescent and adult outcomes that have been associated with poor early reading skills. The study uses data from two cohorts of students drawn from 11 schools in the spring of 1997 and the spring of 1998. The researchers sent letters home to the parents of 723 students that teachers identified as being in the lowest 20% of readers in their classroom. Of these, 295 students were screened using standardized reading and IQ tests. To be eligible for the study, students had to obtain a standard score below 90 on either the Word Identification or the Word Attack subtest of the Woodcock Reading Mastery Tests, obtain a standard score below 90 on a composite of these two subtests, and have a Verbal IQ of at least 80. After screening and balancing for gender, 89 students were randomly assigned to treatment or control (48 to treatment and 41 to control). The study also contained a neuroimaging component that required an additional health screening post randomization; thus the researchers contacted parents again to gain consent for both the neuroimaging and tutoring aspects of the experiment. This resulted in a final sample of 37 students in treatment and 32 students in control. Balance tests

<sup>23</sup> We add to the [Fryer and Dobbie \(2013\)](#) definition because not all studies report days and group size.

revealed no significant differences on observables between the final set of treatment and controls students at baseline.

Treatment students received one-on-one tutoring instruction for 50 min a day, 5 days a week, from September to June. This resulted in the average treatment student attending 126 sessions or 105 h of tutoring. This instruction replaced the typical remedial instruction that the schools offered and that the control students participated in. The instruction was carried out by 12 tutors who were certified in reading or special education. Prior to the intervention, each tutor received 45 h of training on early childhood interventions, early reading acquisition, and teaching strategies. Additionally, tutors received 2 h of training each month for the duration of the experiment. The instruction focused on developing fluency and comprehension strategies, and teaching students to read for pleasure. To do this, tutors incorporated a five-step plan that was featured in previous published studies into each session (Blachman, 1987; Blachman et al., 1999). In over 90% of classroom observations and audiotapes of the tutor sessions, tutors included all five steps of the instruction. Control students continued business as usual—9 control students received no remedial instruction outside of their reading class and the rest participated in small-group tutoring that met for 3–5 times a week. On average, control students that received remedial instruction attended 104 sessions and received 77 h of additional instruction.

To test the impact of treatment, Blachman et al. (2004) administered a battery of tests pretreatment, immediately following treatment, and one year after treatment. The test battery included the Woodcock Reading Mastery Tests—Revised (WRMT), Gray Oral Reading Tests—Third Edition (GORT), Wide Range Achievement Test 3—Spelling (WRAT), and the Calculation and Applied Problems subtests from the Woodcock—Johnson Psycho-Educational Battery—Revised (WJ-R). In addition, the researchers administered subtests of the nonnormed Comprehensive Test of Phonological Processes (CTOPP) 4 times during the treatment year and 4 times during the follow-up year. At posttest, the researchers found large and statistically significant impacts on all standardized reading measures. These impacts ranged from  $0.55\sigma$  on the comprehension subtest of the GORT to  $1.69\sigma$  on the WRMT basic skills cluster. Furthermore, 6 of these 8 impacts were still large and significant a year after the completion of the experiment (the two insignificant impacts were  $0.30\sigma$  and  $0.24\sigma$  on the GORT accuracy and GORT comprehension subtests, respectively). The authors saw similar reading results for the nonstandardized subtests from the CTOPP. As expected, treatment had no significant impacts on standardized measures of mathematics. If anything, at posttest, the math results suggest negative impacts with effect sizes of  $-0.33\sigma$  and  $-0.37\sigma$  on the WJ-R calculations and applied problems subtests, respectively.

The findings overall suggest that one-on-one high dosage tutoring with research-proven instruction can increase the growth rates of low-ability students. Although treatment and control students have statistically indistinguishable growth rates in the follow-up year, the large impact on reading scores from one year of treatment remains.

Another randomized study that investigates the impacts of high-dosage tutoring on low-ability students is Cook et al. (2014). In this study, we implemented an academic and behavioral intervention for 106 male 9th and 10th grade students from a public school in the south side of Chicago. Over 90% of the sample was both black and eligible for free or reduced-price lunch. The intervention consisted of providing students with nonacademic supports that teach the students social cognitive skills through the “Becoming a Man” (BAM) program while also providing students intensive individualized tutoring. The BAM program used principles of cognitive behavioral therapy to deliver a curriculum that focused on values education. The program sought to develop specific social or social cognitive skills such as generating new solutions to problems, learning new ways to behave, and identifying consequences ahead of time. BAM was conducted in small groups that met once a week for 1 h each time. Over the course of the year, students had the chance to participate in 27 different group sessions and typically had to skip an academic class to participate. For the academic portion of the intervention, students met in groups of two with a math tutor 1 h a day, everyday. The tutors were hired following the methodology of Match Corps and were paid \$16,000 plus benefits for the 9-month academic year.<sup>24</sup> Control students were not eligible to participate in BAM or the intensive tutoring but could participate in other academic supports available at the high school.

Students were selected to participate in the study based on an academic risk index that was a function of the number of prior-year course failures, unexcused absences, and being previously held back. The 106 male 9th and 10th grade students with the highest risk index score were then randomly assigned to three groups: Control ( $N = 34$ ), BAM only ( $N = 24$ ), and BAM plus high-dosage tutoring ( $N = 48$ ). To investigate the impact of assignment to one of these two treatment arms, we obtained student-level records from Chicago Public Schools that contained demographic information and scores from the EXPLORE and PLAN tests for the year prior to and year of the intervention.

We found that the ITT effect of assigning students to either one of the treatment arms was large and statistically significant for math achievement and math GPA. Assignment to either treatment arm increased math achievement by  $0.51\sigma$  and increased math GPA by 0.425 grade points on a 4-point scale. We found no spillovers to reading achievement and no significant impacts on discipline incidents or number of days suspended. However, treatment students were absent 10.272 fewer days throughout the school

<sup>24</sup> Match Corps is an AmeriCorps program in which members spend a year attempting to close the achievement gap by tutoring small groups of students in the various Match Charter Schools in Boston. Match Corps seeks to employ tutors who are dedicated to constant improvement, who possess strong communication and writing skills, and who are committed to spending a year working with children. Adopting their hiring best practices, tutors in Chicago were required to pass a math assessment, conduct a mock tutorial session with actual high school students, and interview with the principal or principal's designee.

year. Mostly due to the relatively modest size of our sample, when separated by treatment arm, we found no significant difference between the impacts of the two groups. Summarizing, the metacoefficient on high-dosage tutoring is  $0.309\sigma$  (0.106) for math achievement and  $0.229\sigma$  (0.033) for reading achievement. Indeed, 54.3% of coefficients demonstrate statistically significant positive treatment effects; 0% yield statistically significant negative effects. Surprisingly, the fraction of statistically positive treatments is larger than early childhood interventions.

**3.3.1.3.2 Low-dosage tutoring** The Early Start to Emancipation Preparation (ESTEP)-Tutoring program of Los Angeles County was created in 1998. The program targets foster children aged 14 to 15 who are 3 or more years behind in math or reading ability. ESTEP-Tutoring aims to improve the math and reading skills of these students and encourage them to take advantage of educational resources of which they may have been previously unaware. Tutoring is provided in the home of the students by college student tutors drawn from the surrounding 12 community colleges. Tutors are trained to teach the students in math, reading, and spelling and are provided with curriculum materials that fit a student's skill level. In addition to tutoring, the program hopes to foster a mentorship relationship between the tutor and the student. Once assigned to the program, each student is eligible for 50 h of tutoring, and tutors are allotted additional time for preparation, mentoring, or other activities. [Courtney et al. \(2008\)](#) take advantage of the high demand of ESTEP-Tutoring and conduct an evaluation of the program using its oversubscribed application pool.

For the study, all students referred to the program were screened to ensure their math or reading ability was indeed 3 years behind grade level. Eligible students were then randomly assigned to a group that could participate in ESTEP-Tutoring or a control group that could not. This resulted in a study sample of 445 students, 246 assigned to treatment and 219 assigned to control. On average, approximately 4 months passed between assignment and a student's first meeting with a tutor. Throughout the 2 years of the study researchers found that 61.8% of treatment students eventually participated in ESTEP-Tutoring and the average treatment student received 18 h of math tutoring and 17 h of reading tutoring. The relatively low take-up rate is attributed to the high mobility of foster children and the length of time that passed between assignment and receipt of tutoring. By the time tutors attempted to deliver the first tutoring session, a majority of the nonparticipants were no longer in the foster home listed on their application. Once the tutoring program was initiated, students were eligible for 50 h of tutoring delivered through 2 h sessions twice a week.

To investigate the impact of ESTEP-Tutoring on these students, [Courtney et al. \(2008\)](#) conducted three interviews over the 2 years after randomization (baseline, 1 year out, and 2 years out). At each of these interviews, the researchers administered the letter-word identification, calculation, and passage comprehension subtests of the

Woodcock–Johnson Tests of Achievement III as well as a student survey. The survey combined questions from The Midwest Evaluation of Adult Functioning of Former Foster Youth, The National Survey of Child Adolescent Well-Being, the National Longitudinal Survey of Youth, and the National Longitudinal Survey of Adolescent Health. The survey collected data on demographics, prior experiences in care, prior victimization, relationships, social support, employment, education, health behaviors, and physical health.

Courtney et al. (2008) found evidence of a first-stage impact in that treatment students were more likely to report having been tutored at home. However, control students were more likely to report that they had received tutoring at school and the total number of tutoring hours reported by treatment and control students were not statistically different. The authors limit their impact analysis to the second follow-up interview (2 years after random assignment) due to the fact that participation in ESTEP was still ongoing for many students 1 year after random assignment and they find no evidence of impacts on any outcome measure. The difference between control and treatment groups on Woodcock–Johnson achievement scores, school grades, educational attainment, and school behavior are all statistically indistinguishable from zero. On putting all low-dosage tutoring experiments together, the metacoeficient on low-dosage tutoring is  $0.015\sigma$  (0.013) for math achievement and  $0.015\sigma$  (0.015) for reading achievement.

### **3.3.2 Teacher-based interventions**

Great teachers matter. A one-standard deviation improvement in teacher quality translates into annual student achievement gains of  $0.15\sigma$  to  $0.24\sigma$  in math and  $0.15\sigma$  to  $0.20\sigma$  in reading (Rockoff, 2004; Rivkin et al., 2005; Aaronson et al., 2007; Kane and Staiger, 2008). These effects are comparable to reducing class size by about one-third (Krueger, 1999). Using quasiexperimental methods, Chetty et al. (2011) estimate that a one-standard deviation increase in teacher quality in a single grade increases earnings by about 1% per year; students assigned to these better teachers are also more likely to attend college and save for retirement, and less likely to have children when teenagers.

How to select or produce great teachers is one of the most important open questions in human capital research. Observable characteristics such as college-entrance test scores, grade point averages, or major choice are not highly correlated with teacher value-added on standardized test scores (Aaronson et al., 2007; Rivkin et al., 2005; Kane and Staiger, 2008; Rockoff et al., 2012). And, many programs that aim to make teachers more effective have shown little impact on teacher quality (see e.g., Boyd et al., 2007 for a review). Some argue that these two facts, coupled with the inherent costs of removing low performing teachers due to collective bargaining agreements along with increased job market opportunities for women, contributes to

the fact that teacher quality and aptitude has declined significantly in the past 40 years ([Corcoran et al., 2004; Hoxby and Leigh, 2004](#)).

We group the set of teacher-based random assignment studies into three subcategories: increasing teacher supply, providing teachers incentives, or increasing human capital through professional development.

### 3.3.2.1 Increasing teacher supply

Perhaps the most obvious way to increase teacher supply is to lower the barriers into the teaching profession by allowing alternative routes for teachers to obtain necessary certifications. Due to the teacher shortages and the No Child Left Behind Act, which required every classroom to be staffed with a certified teacher or a teacher actively pursuing a certification through an approved program, there has been an increase in teachers who enter teaching through alternative paths. Traditionally, teachers have completed all of their certification requirements at an accredited university or program before starting to teach in a classroom. In comparison, alternatively certified (AC) teachers start teaching before completing their requirements and earn their certification while teaching. Well-known examples of AC programs include Teach for America (TFA) and the New York City Teaching Fellows program. Both of these programs attract extremely qualified uncertified individuals and place them in schools that are in dire need of good teachers. The potential benefits and advantages of these different routes to certification have been debated by many. For example, some argue that the coursework required for traditionally certified (TC) teachers is an unnecessary burden that discourages some from pursuing teaching and AC programs are a way to circumvent that. In contrast, others argue that without that coursework, AC teachers enter classrooms underprepared and will be less effective.

To better understand the effectiveness of AC teachers relative to TC teachers, [Constantine et al. \(2009\)](#) conducted a randomized study in elementary schools around the nation in the 2004–2005 and 2005–2006 school-years. Their study included 63 schools from 20 districts in seven states across the nation. Within these schools, 2610 K–5 students were randomly assigned to be taught by an AC teacher or a TC teacher for one school-year. Schools were only allowed to participate if they had at least one eligible AC teacher and one eligible TC teacher in the same grade. In order for a teacher to be eligible to participate, teachers had to be, relatively speaking, novices, had to teach in a regular classroom, and had to deliver both reading and math instruction to all their students. Researchers collected data on student achievement by administering the math and reading sections of the California Achievement Test, 5th Edition (CAT). The researchers also collected data on the classroom practices of teachers through classroom observations and principals' ratings. In addition, all teachers completed a survey in the spring that collected information on teachers' professional and personal backgrounds, experience in the school as a full-time teacher, and SAT/ACT scores.

Finally, they also collected data on the details of each program a teacher attended for certification/alternative placement.

Constantine et al. (2009) found that students of AC teachers did not perform statistically differently than students of TC teachers. Furthermore, there were no statistically significant differences when comparing low grade-level (K-1) teachers to high grade-level (2–5) teachers or low-experience teachers to high-experience teachers. When exploring heterogeneous effect sizes across amount of coursework teachers were required to do while teaching, there is some evidence that AC teachers who had high levels of coursework had negative impacts on student achievement. Similarly, there is no statistically significant difference in classroom observation scores between AC and TC teachers. However, when restricting the sample to teachers that had high levels of coursework, there is evidence that AC teachers' classroom practices were worse than TC teachers' practices.

In addition to the experimental results, Constantine et al. (2009) also present nonexperimental results that explore the relationship between teacher characteristics and program details with the impacts on students' achievement. Overall, they found that teacher characteristics and training experiences only explained 5% of the variation in effects on math test scores and 1% of the variation in effects on reading test scores. The only significant correlations they found were that AC teachers with master's degrees were less effective in improving student achievement in reading than TC teachers without a master's degree and that students in classrooms taught by AC teachers who were taking coursework towards a degree or certification did worse in reading than students taught by TC teachers who weren't taking coursework.

Some argue that schools do not just need access to more teachers, but specifically need access to different and potentially better talent pools. Proponents of this argument often point to successful foreign education systems, such as Hong Kong or Finland, that draw their teachers from the uppermost ranks of their universities (Tucker, 2011). In contrast, it is a well-documented fact that the talent pool of American teachers has been declining since 1960 (see Corcoran et al., 2004). Hoxby and Leigh (2004) attribute a large part of this decline to opportunities outside of teaching drawing high-aptitude women from the profession. Over the past couple decades, we have seen an increase of programs designed to combat this decline and get more and better college students to enter into teaching. One such program is Teach for America.

Teach for America, a nonprofit organization that recruits recent college graduates to teach for 2 years in low-income communities, is one of the United States most prominent service programs. Based on founder Wendy Kopp's undergraduate thesis at Princeton University, TFA's mission is to create a movement that will eliminate educational inequity by enlisting our nation's most promising future leaders as teachers. In 1990, TFA's first year in operation, Kopp raised \$2.5 million and attracted 2500 applicants for 500 teaching slots in New York, North Carolina, Louisiana, Georgia, and Los Angeles.

Since its founding, TFA corps members have taught more than three million students. Today, there are 8200 TFA corps members in 125 “high-need” districts across the country, including 13 of the 20 districts with the lowest graduation rates. Roughly 80% of the students reached by TFA qualify for free or reduced-price lunch and more than 90% are black or Hispanic.

Entry into TFA is highly competitive; in 2010, more than 46,000 individuals applied for just over 4000 spots. Twelve percent of all Ivy League seniors applied. In its recruitment efforts, TFA focuses on individuals who possess strong academic records and leadership capabilities, regardless of whether or not they have had prior exposure to teaching. To apply, candidates are required to complete an online application, which includes a letter of intent and a resume. After a telephonic interview, the most promising applicants are invited to participate in an in-person interview, which includes a sample teaching lesson, a group discussion, a written exercise, and a personal interview. Applicants who are invited to interview are also required to provide transcripts, obtain two online recommendations, and provide one additional reference.

Using information collected through the application and interview, TFA bases their candidate selection on a model that accounts for multiple criteria that they believe are linked to success in the classroom. These criteria include achievement, perseverance, critical thinking, organizational ability, motivational ability, respect for others, and commitment to the TFA mission. TFA conducts ongoing research on their selection criteria, focusing on the link between these criteria and observed single-year gains in student achievement in TFA classrooms.

TFA teachers are required to take part in a 5-week TFA summer institute to prepare them for placement in the classroom at the end of the summer. The TFA summer institute includes courses covering teaching practice, classroom management, diversity, learning theory, literacy development, and leadership. During the institute, groups of participants also take full teaching responsibility for a class of summer school students.

At the time of their interview, applicants submit their subject, grade, and location preferences. TFA works to balance these preferences with the needs and requirements of districts. With respect to location, applicants rank each TFA region as highly preferred, preferred, or less preferred and indicate any special considerations, such as the need to coordinate with a spouse. Over 90% of the TFA applicants accepted are matched to one of their “highly preferred” regions (Glazerman et al., 2006).

TFA also attempts to match applicants to their preferred grade levels and subjects, depending on applicants’ academic backgrounds, district needs, and state and district certification requirements. As requirements vary by region, applicants may not be qualified to teach the same subjects and grade levels in all areas. It is also difficult for school regions to predict the exact openings they will have in the fall, and late changes in subject or grade-level assignments are not uncommon. Predicted effectiveness scores are not used to determine the placement region, grade, or school, and the scores are not available to districts.

TFA corps members are hired to teach in local school districts through alternative routes to certification. Typically, they must take and pass exams required by their districts before they begin teaching and may also be required to take additional courses to meet state certification requirements.

TFA corps members are employed and paid directly by the school districts for which they work, and generally receive the same salaries and health benefits as other first year teachers. Most districts pay a \$1500 per corps member fee to TFA to offset screening and recruiting costs. TFA gives corps members various additional financial benefits, including “education awards” of \$4725 for each year of service that can be used for past or future educational expenses, and transitional grants and no-interest loans to help corps members make it to their first paycheck.

To date, there have been a couple randomized evaluations of the impact of TFA teachers. [Glazerman et al. \(2006\)](#) report findings from a national evaluation of TFA. The experiment involved approximately 100 elementary classrooms from 17 schools drawn from Baltimore, Chicago, Compton, Houston, New Orleans, and the Mississippi Delta. Students were stratified by grade and school and assigned randomly to either a TFA or a nonTFA teacher. At the end of school-year, [Glazerman et al. \(2006\)](#) found that students assigned to TFA teachers score about  $0.12\sigma$  higher in math and  $0.03\sigma$  higher in reading than students assigned to traditionally certified teachers. They found no impacts on other student outcomes such as attendance, promotion, or disciplinary incidents, but TFA teachers were more likely to report problems with student behavior than were their peers.

An even bigger study analyzed by [Clark et al. \(2013\)](#) uses a sample drawn from almost 100 schools across eight states to investigate the effectiveness of middle school math teachers from TFA and a similar program called The New Teacher Project (TNTP). In each participating school, students were randomly assigned to math classrooms taught by a program teacher (TFA or TNTP) or a teacher did not enter teaching through either of these programs. Similar to [Glazerman et al. \(2006\)](#), [Clark et al. \(2013\)](#) find a significant impact of TFA on students’ math test scores. Students assigned to TFA teachers scored  $0.07\sigma$  higher on state standardized testing whereas students assigned to TNTP teachers had test scores that were indistinguishable from students in control classrooms. Note that this study was not designed to investigate the difference between TFA and TNTP teachers. Students were not randomly assigned between TFA and TNTP teachers, so differences between the effectiveness of the teachers could be due to differences in the students they taught, the comparison teachers, or the schools they were in. Indeed, TFA and TNTP teachers included in the study largely taught in different schools and districts. With this in mind, there are still some major differences between the two programs worth noting. TFA requires its teachers to commit to 2 years of teaching under the TFA arrangement whereas TNTP expects their recruits to teach

for many years. Also, TFA recruits heavily from college campuses while TNTP recruits professionals that want to switch careers.

Several other programs similar—in spirit—to TFA are Boston Teaching Residency, Match Teaching Residency, NYC Teaching Fellowships, Inner-City Teaching Corps of Chicago, and Harvard Teaching Fellows. Although these programs all differ in length, training procedures, and credentials earned through the program, they all recruit college graduates with strong academic backgrounds and place them in struggling school districts. To the best of our knowledge, no randomized evaluations exist yet for these programs.

### 3.3.2.2 Teacher incentives

To increase teacher productivity, there is growing enthusiasm among policymakers for initiatives that tie teacher incentives to the achievement of their students. Since 2006, the US Department of Education has provided over \$1 billion to incentive programs through the Teacher Incentive Fund—a program designed specifically to support efforts developing and implementing performance-based compensation systems in schools. At least seven states and many more school districts have implemented teacher incentive programs in an effort to increase student achievement (Fryer, 2013b).

Yet, the empirical evidence on the effectiveness of teacher incentive programs is mixed. In developing countries where the degree of teacher professionalism is extremely low and absenteeism is rampant, field experiments that link pay to teacher performance are associated with substantial improvements in student test scores (Duflo et al., 2012; Glewwe et al., 2010; Muralidharan and Sundararaman, 2011). Conversely, the few field experiments conducted in the United States have had, at best, mixed results.

Theoretically, it is unclear how to design optimal teacher incentives when the objective is to improve student achievement. Much depends on the characteristics of the education production function. If, for instance, the production function is additively separable, then individual incentives may dominate group incentives, as the latter encourages free riding. If, however, the production function has important complementarities between teachers in the production of student achievement, group incentives may be more effective at increasing achievement (Baker, 2002).

**3.3.2.2.1 Group incentives** In the 2007–2008 through the 2009–2010 school year, the United Federation of Teachers (UFT) and the New York City Department of Education (DOE) implemented a teacher incentive program in over 200 high-need schools, distributing a total of roughly \$75 million to over 20,000 teachers.<sup>25</sup> The experiment was a randomized school-based trial. Each participating school could earn \$3000

<sup>25</sup> The details of the program were negotiated by Chancellor Joel Klein and Randi Weingarten, along with their staffs. At the time of the negotiation, I was serving as an advisor to Chancellor Klein and convinced both parties to agree to include random assignment to ensure a proper evaluation.

for every UFT-represented staff member if the school met the annual performance target set by the DOE based on school report cards, which the school could distribute at its own discretion. Each participating school was given \$1500 per UFT staff member if it met at least 75% of the target but not the full target. Note that the average New York City public school has roughly 60 teachers; this implies a transfer of \$180,000 to schools on average if they met their annual targets and a transfer of \$90,000 if they met at least 75%, but not the full target. In elementary and middle schools, school report card scores hinge on student performance and progress on state assessments, student attendance, and learning environment survey results. High schools are evaluated similarly, with graduation rates, Regents exams, and credits earned replacing state assessment results as proxies for performance and progress.

An important feature of the experiment is that schools had discretion over their incentive plans. As mentioned above, if a participating school met all of the annual targets, it received a lump sum equivalent to \$3000 per full-time unionized teacher. Each school had the power to decide whether all of the rewards would be given to a small subset of teachers with the highest value-added, whether the winners of the rewards would be decided by lottery, or virtually anything in-between. The only restriction was that schools were not allowed to distribute rewards based on seniority.

An overwhelming majority of the schools decided on a group incentive scheme that varied the individual bonus amount only by the position held in the school. This could be because teachers have superior knowledge of education production and believe the production function to have important complementarities, because they feared retribution from other teachers if they supported individual rewards, or simply because this was as close to pay based on seniority (the UFT's official view as to why schools typically settled on this scheme) as they could do.

The results from this incentive experiment are informative. Providing incentives to teachers based on a school's performance on metrics involving student achievement, improvement, and the learning environment did not increase student achievement in any statistically meaningful way. If anything, student achievement declined. ITT estimates yield treatment effects of  $-0.018\sigma$  (0.024) in mathematics and  $-0.014\sigma$  (0.020) in reading for elementary schools, and  $-0.046\sigma$  (0.018) in math and  $-0.030\sigma$  (0.011) in reading for middle schools, *per year*. Thus, if an elementary school student attended schools that implemented the teacher incentive program for 3 years, her test scores would decline by  $-0.054\sigma$  in math and by  $-0.042\sigma$  in reading, neither of which is statistically significant. For middle school students, however, the negative impacts are more sizeable:  $0.138\sigma$  in math and  $-0.090\sigma$  in reading over a three-year period.

Consistent with Fryer (2013b), Springer et al. (2012) evaluated another group incentive experiment that took place in the Round Rock Independent School District in Texas. The study used random assignment to investigate the impacts of a program that awarded teams of middle school teachers bonuses based on their collective contribution

to students' test score gains. Two years after the initial randomization, Springer et al. (2012) found no significant impacts on the attitudes and practices of teachers or on the academic achievement of students.

**3.3.2.2 Individual incentives** Springer et al. (2010) evaluated Tennessee's POINT program—a 3-year pilot initiative on teacher incentives conducted in the Metropolitan Nashville School System from the 2006–07 school year through the 2008–09 school year. 296 middle school mathematics teachers who volunteered to participate in the program were randomly assigned to the treatment or the control group, and those assigned to the treatment group could earn up to \$15,000 as a bonus if their students made gains in state mathematics test scores equivalent to the 95th percentile in the district. They were awarded \$5000 and \$10,000 if their students made gains equivalent to the 80th and the 90th percentiles, respectively. Springer et al. (2010) found there was no significant treatment effect on student achievement and on measures of teachers' response such as teaching practices.

In an important observation, Neal (2011) discusses how group incentives (e.g., Fryer, 2013b; Springer et al., 2012) or sufficiently obtuse (e.g., Springer et al., 2010) pay schemes lead to problems when trying to calculate the incentive effect at the individual teacher level and could be the reason these experiments observed little to no incentive effects. For instance, calculating the expected value of a one standard deviation increase in teacher effort when the incentive scheme depends on where a teacher lies in the overall district distribution is a nontrivial calculation for an econometrician with loads of data and sophisticated techniques. It would be exceedingly difficult for a teacher to perform this calculation and understand how their efforts could translate into rewards. To circumvent this and competition issues between teachers, Barlevy and Neal (2012) develop a "pay for percentile" method that rewards teachers according to how highly their students' test score improvement ranks among other students from other schools with similar baseline achievement and demographic characteristics.

Although not fully using the method recommended by Barlevy and Neal (2012), Glazerman et al. (2009) present results from a randomized experiment that ties individual teacher incentives to value-added measures. This incentive scheme is more in line with the insights in Neal (2011) than Fryer (2013b) or Springer et al. (2010, 2012).

In 2007, Chicago Public Schools implemented its own version of the national Teacher Advancement Program (TAP). The national version of TAP was developed in the late 1990s by the Milken Family Foundation as an incentive program to increase teacher quality. Teachers could earn extra pay by being promoted to Mentor or Lead Teacher and receive annual performance bonuses based on their value-added and classroom observations. Chicago adopted this model with some minor alterations. For example, the Chicago TAP added principal bonuses tied to implementation

benchmarks and school-wide value-added. Teacher incentives had an expected payout of \$2000 per teacher and teachers could earn an additional \$7000 by becoming a Mentor or an additional \$15,000 by becoming a Lead Teacher. As Mentors, teachers were expected to provide ongoing classroom support to other teachers in the school. Lead Teachers served on the leadership team responsible for implementing TAP, analyzing student data, and developing achievement plans. In addition, Mentors and Lead Teachers conducted weekly group meetings to foster collaboration between teachers and provide additional professional development.

[Glazerman et al. \(2009\)](#) conducted a randomized evaluation of the first year of Chicago TAP. Of the 16 K–8 schools that volunteered to participate in the program, eight were randomly assigned to start treatment in the 2007–2008 school year and the other eight would delay the start of the program until the 2008–2009 school-year. [Glazerman et al. \(2009\)](#) compared the outcomes for teachers and students in schools randomly assigned to the two groups for the 2007–2008 school year to determine causal impacts of exposure to 1 year of Chicago TAP. For their analysis, the researchers collected student achievement data and teachers' classroom assignments from Chicago Public Schools as well as administered surveys to teachers and principals to collect important information that was not present in the administrative data.

The evaluation suggests that TAP increased retention in treatment schools. Teachers in TAP schools had a retention rate of 87.9% while teachers in control schools had a retention rate of 82.8%, a statistically significant difference. However, teacher satisfaction and teachers' positive attitudes toward their principals were not statistically different between TAP and control schools.

More importantly, the introduction of TAP did not produce any measurable impacts on student standardized test scores. The effect size for reading was  $-0.04\sigma$  (0.05) and the effect size for math was  $-0.04\sigma$  (0.06). The test score impacts were insignificant across all grade levels and were robust to various sensitivity analyses.

**3.3.2.2.3 Enhancing the efficacy of teacher incentives through framing** During the 2010–2011 and the 2011–2012 school years, [Fryer et al. \(2015a\)](#) conducted an experiment in nine schools in Chicago Heights, IL. At the beginning of each school year, teachers were randomly selected to participate in a pay-for-performance program. Among those who were selected, the timing and framing of the reward payment varied. One set of teachers whom we label the “Gain” treatment received “traditional” financial incentives in the form of bonuses at the end of the year linked to student achievement.<sup>26</sup> Other teachers—the “Loss” treatment—were

<sup>26</sup> Note that although math, reading, and science test scores were incentivized (the latter only for 4th and 7th grade science teachers), the main analysis of the paper focuses on math achievement due to most students having multiple reading teachers and the science sample being so small.

given a lump sum payment at the beginning of the school year and informed that they would have to return some or all of it if their students did not meet performance targets. Teachers in the “Gain” and “Loss” groups with the same performance received the same final bonus. Within the “Loss” and “Gain” groups, we additionally tested whether there are heterogeneous effects for individual teacher rewards compared to awarding incentives to teams of teachers.

In all groups, performance was incentivized according to the “pay for percentile” method developed by [Barlevy and Neal \(2012\)](#), in which teachers are rewarded according to how highly their students’ test score improvement ranks among peers from other schools with similar baseline achievement and demographic characteristics. As [Neal \(2011\)](#) describes, pay for percentile schemes separate incentives and performance measurements for teachers since this method only uses information on relative ranks of the students. Thus, motivation for teachers to engage in behaviors (e.g., coaching or cheating) that would contaminate performance measures of the students is minimized.

The first year ITT results of our experiment are consistent with over 3 decades of psychological and economic research on the power of framing to motivate individual behavior, though other models may also be consistent with the data. Students who were assigned to teachers in the “Loss” treatment show large and statistically significant gains in year one math test scores [ $0.455\sigma$  (0.097)]. Teacher incentives that are framed as gains demonstrate less success. In the first year of the experiment, students in the “Gain” treatment increased their math test scores  $0.245\sigma$  (0.094). Importantly, the difference between the “Loss” and “Gain” treatments in math improvement is statistically significant at conventional levels. More generally, these results support the view in [Barlevy and Neal \(2012\)](#), and [Neal \(2011\)](#) that properly designed incentives can have significant effects.

Interestingly, when looking at the sample of all students, we find little evidence of treatment effects in the second year of the experiment. The ITT estimates for “Loss” are  $0.087\sigma$  (0.088) and for “Gain” are  $0.115\sigma$  (0.109). The pooled estimates for both years of the experiment are  $0.210\sigma$  (0.069) and  $0.116\sigma$  (0.075) for “Loss” and “Gain,” respectively. The difference between the “Loss” and “Gain” treatments for the pooled estimates has a *p*-value of 0.099.

Although incentivizing teachers had differential impacts across both years when looking at the entire sample, we found that kindergartners had large gains in both years of the experiment regardless of whether their teachers were in the “Loss” or “Gain” group. In the first year of the experiment, the ITT estimates for kindergarten students were  $0.796\sigma$  (0.209) for “Loss” and  $0.376\sigma$  (0.168) for “Gain.” In the second year, the estimates were  $0.574\sigma$  (0.176) and  $0.714\sigma$  (0.144) for “Loss” and “Gain” respectively. Therefore, the pooled effect size for both years and both treatments was  $0.568\sigma$  (0.121) for kindergarten math scores.

**3.3.2.2.4 Talent transfers** In America, inexperienced teachers are more likely to be assigned to high-minority and high-poverty classrooms (Feng, 2010). As a result, novice teachers are taking on tougher school assignments, teaching multiple grades, and teaching out-of-field classes (Donaldson and Johnson, 2010). To counteract this trend, several school districts—such as Houston ISD—provide effective teachers incentives to teach in the most troubled schools. The theory is that the marginal return for an additional effective teacher in a well-functioning school is less than the marginal return of that teacher in a less well-functioning school. Good teachers have the potential to change the culture of a school and provide effective pedagogical tools and mentorship to struggling colleagues. If true, providing incentives for talented teachers to teach in troubled schools will increase total productivity.

Glazerman et al. (2013) used a randomized experiment in 10 districts across the nation to investigate the impact of filling vacancies with high-achieving teachers through the Talent Transfer Initiative (TTI). In each district, the TTI offered teachers with consistently high value-added (ranking in the top 20% within their subject and grade) \$20,000, paid over 2 years, to teach at low-achieving schools selected through a random process. Principals of schools with low average test scores volunteered to fill vacancies at their school using the TTI. Schools that volunteered were matched based on the grade-level and subject of the vacancy as well as school demographics. Teacher teams (teachers grouped by grade and subject) within each block of schools that had at least one vacancy were then randomly assigned to treatment or control. Teacher teams assigned to treatment status were eligible to fill their vacancy with a TTI teacher and vacancies in control teacher teams were filled using the typical process of the given school. Note that high-performing teachers were not randomly assigned to these vacancies. After a teacher team is assigned to treatment, TTI teachers must interview for the position, principals must extend an offer to a TTI teacher, and then a TTI teacher must accept and voluntarily move to fill this vacancy. To receive the full financial incentive, high-performing teachers must remain in the low-achieving school for a full 2 years.

Glazerman et al. (2013) investigated the impact of teacher teams being eligible to fill their vacancies using the TTI. Across the 10 districts included in the study, 165 teacher teams from 114 schools were randomly assigned to treatment ( $N = 85$ ) or control ( $N = 80$ ). The transfer incentive was able to successfully attract high-achieving teachers. Eighty-eight percent of treatment vacancies were filled with teachers through the TTI. To achieve this high rate of transfer, over 1500 high-achieving teachers were invited to participate in TTI. The teachers hired to fill treatment vacancies were significantly more experienced than teachers hired for control spots. Treatment teachers had on average 4 years more experience and were 11 percentage points more likely to have a National Board Certification than control teachers ( $CM = 9\%$ ). Interestingly, there was evidence that principals reacted to the hiring of a TTI teacher by reallocating weak teachers to be

in the same team as the incoming TTI teacher. Teachers from elsewhere in the school that joined a treatment teacher team after the hiring of a TTI teacher had 5 years less experience than teachers that moved into a control teacher team. In addition, treatment teachers were more likely to provide mentoring to their peers (15 compared to 5% of the control teachers) and were less likely to receive mentoring (39 compared to 59% of control teachers).

[Glazerman et al. \(2013\)](#) found that the above first-stage and intermediate impacts translated into large effects on student achievement tests for elementary schools but that there were no significant impacts on the achievement of middle school students. TTI eligible elementary classrooms increased students' math scores by  $0.18\sigma$  and students' reading scores by  $0.10\sigma$  in the first year after randomization. In the second year, the cumulative impacts for treatment students were  $0.22\sigma$  and  $0.25\sigma$  for math and reading test scores, respectively. Although treatment elementary teachers had large impacts on their students, there were no spillover effects for other teachers in their team. Elementary student achievement outcomes were not significantly different between students assigned to other teachers in the treatment team and students assigned to other teachers in the control team.

Finally, there was evidence that the TTI was effective at keeping these high-performing teachers in the low-performing schools. At the halfway point of the program, retention rates were higher for teachers that filled TTI vacancies. Treatment teachers were 23 percentage points more likely to remain in a school after the first year than their control counterparts (93% compared to 70%). In addition, retention rates after the completion of the second year of the experiment were not statistically significant. Approximately 60% of treatment teachers returned to the low-achieving school for a third, nonincentivized school year. In comparison, a statistically indistinguishable 51% of control teachers remained in the fall of the third school year.

The metacoefficient on teacher incentives is  $0.022\sigma$  (0.022) for math achievement and  $-0.006\sigma$  (0.012) for reading achievement. Yet, that number seems particularly misleading in this context as many of the schemes were quite ad hoc and inconsistent with economic theory. More experiments are needed before one can better hazard a guess on the efficacy of teacher incentives. Future randomized trials ought to take the insights in [Barlevy and Neal \(2012\)](#) and [Neal \(2011\)](#) seriously when designing teacher incentive schemes.

### 3.3.2.3 Teacher professional development

**3.3.2.3.1 General professional development** The Gates Foundation states that the American education system spends \$18 billion annually on professional development ([Bill and Melinda Gates Foundation, 2014](#)). For 2014, Title II of the Elementary and Secondary Education Act, a program mostly devoted to professional development, was appropriated \$2.3 billion ([US Department of Education, \(2014\)](#)). More than \$450

million (approximately half) of the Department of Education’s Investing in Innovation (i3) grant money funded professional development programs from 2010 to 2012 ([US Government Accountability Office, 2014](#)). A new report released by TNTP estimates that three large public districts included in their study spent nearly \$18,000 per teacher per year for professional development ([TNTP, 2015](#)).

Professional development (PD) is viewed as a vital tool to increase teachers’ human capital and improve school effectiveness ([Hill, 2007](#)). However, experts have expressed concern that teachers are not receiving enough professional development to have meaningful impacts on teachers’ practices and that the little professional development they receive does not focus enough on subject-matter knowledge ([Cohen and Hill, 2001](#); [Fletcher and Lyon, 1998](#); [Foorman and Moats, 2004](#); [Garet et al., 2001](#)).<sup>27</sup> Another often articulated concern is that professional development tends to be one-time workshops scheduled on “professional development days” or in the summer months with little relevant follow-up ([Joyce and Showers, 1988](#); [Parsad et al., 2001](#); [Loucks-Horsley et al., 1998](#)).

The US Department of Education commissioned two PD interventions to provide states and districts with further information of the potential of PD programs to improve reading instruction ([Garet et al., 2008](#)). The first intervention provided 2nd grade teachers with a year-long research-based institute series and the second intervention provided the same institute series plus in-school coaching. [Garet et al. \(2008\)](#) presented results from the randomized evaluation of these two interventions. In their study, 90 schools across six districts from four states were randomly assigned to one of the two treatment groups or a control group such that each district had an equal number of elementary schools allocated to the three groups. [Garet et al. \(2008\)](#) collected data on teacher knowledge, teacher practices, and student achievement at the completion of the intervention and 2 years after randomization as a follow-up.

On average, teachers in the first intervention reported attending 39 h of PD and teachers in the second intervention reported attending 47 h of PD. In comparison, control teachers only reported attending 13 h of PD. [Garet et al. \(2008\)](#) found that this exposure to PD led to significant impacts on teachers’ knowledge and practices for both groups. Both interventions had positive impacts on 2nd grade teachers’ knowledge of early reading content and instructional knowledge at posttest and a year after the PD programs had completed. For both years, teachers in both interventions used explicit instruction to a much greater extent than teachers assigned to control. However, there was no significant difference in the amount of independent student activity incorporated into the classroom and the use of differential instruction between either of the two treatment

<sup>27</sup> A national study revealed that over 80% of elementary and secondary teachers reported participating in 24 h or less of professional development over the 2005–2006 school year and summer of 2006 ([US Department of Education, 2009](#)).

groups and control teachers. Although there were large impacts on teacher knowledge and practices, there was no evidence that these changes had an impact on students' test scores. Test scores from the implementation year and the follow-up year revealed no statistically significant impacts on standardized math and reading outcomes.

Another widely used PD program is Classroom Assessment of Student Learning (CASL). The program set consists of a primary text, DVDs, ancillary books, and an implementation handbook. CASL is designed to be a self-executing PD program where teachers learn from the textbook and use CASL assessments to better understand their own and their students' progress. The program mostly emphasizes formative assessments, but also includes lessons on how to utilize other forms of classroom assessments such as standardized test scores. The program is typically implemented via teacher learning teams, in which teachers can discuss and receive feedback from other teachers who are also using the program.

To better understand the effects of CASL on students' achievement, motivation to learn, and teachers' classroom assessment practices, [Randel et al. \(2011\)](#) conducted a large randomized experiment. Due to regional needs, they decided to focus the study in mathematics classrooms.

Almost 70 schools from 32 districts from across Colorado participated in the study. Schools volunteered to participate and were eligible if they were large enough to have at least one 4th grade and one 5th grade teacher. The 67 eligible schools were then randomly assigned to a treatment group ( $N = 33$ ) and a control group ( $N = 34$ ). In November of 2007, treatment schools received one set of CASL PD materials for each math teacher in fourth or 5th grade. In total, there were 178 such teachers in treatment schools and 231 teachers in control schools. Treatment teachers participated in an introductory video conference with the author of CASL and had access to a facilitator who had received training in the CASL program, but other than this, the experiment was completely hands off. The teachers were asked to use the PD naturally without any input or requirements from the research team. The 2007–2008 school year was used as a training year during which teachers studied the CASL material and started integrating CASL practices into their classrooms. The 2008–2009 year was the actual intervention year. Fidelity of treatment was assessed using self-reported logs that 90% of teachers returned to the research team. To combat the alternative hypothesis that any impact was just the result of the intervention schools having more resources, control schools were given \$1000.

To assess the impact of the intervention on student and teacher outcomes, [Randel et al. \(2011\)](#) collected administrative and survey data from both the training and implementation year. The administrative data came directly from the Colorado Department of Education and contained state achievement test scores and student demographics. To quantify students' motivation to learn, researchers administered a student survey. Further, teachers' knowledge of classroom assessments, their classroom assessment practices, and

teachers' involvement of students in assessments were all measured through self-reported teacher surveys.

Randel et al. (2011) found evidence that treatment had a significant impact on teacher knowledge of class assessments. Intervention teachers on average answered 2.78 questions more ( $0.42\sigma$ ) correctly on a 60-item test about teachers' knowledge of classroom assessments. However, there was no evidence that this knowledge influenced their classroom practices. There were no significant differences in classroom practices and the extent to which they involved their students in formative assessments. Intervention teachers were given an average rating of 1.61 for classroom assessment and control teachers had an average rating of 1.60 (where 1 represents low quality and 4 represents high quality). For student involvement, intervention teachers self-reported average score was a 0.39 and control teachers' average score was 0.34 (where 1 indicates that all students were involved in formative assessments everyday and 0 represents no students were involved).

As one would suspect from the similar practices of control and treatment teachers, average student mathematics achievement was not statistically different between treatment and control students. The adjusted mean scale score for students in a treatment classroom was 502.49(2.52) and the mean scale score for students in control classrooms was 501.91(2.44). The impacts remain statistically insignificant when looking at effect sizes by grade level.

The metacoeficient for general PD is  $0.019\sigma$  (0.024) for math achievement and  $0.022\sigma$  (0.023) for reading achievement. In fact, there is not a single study with significant annual pooled impacts. Ironically—and perhaps sadly—Erik Hanushek argues school districts are overspending on ineffective and unmanaged professional development and these districts refuse to veer away from practices that fail time and time again (Layton, 2015).

**3.3.2.3.2 “Managed” professional development** Another form of PD is one that has precise training and curriculum materials that schools and districts can implement to increase teacher effectiveness. These programs are significantly more prescriptive. They do not abstractly discuss issues such as “classroom management” or endeavor to increase “rigor.” Consider two well known examples of this approach to professional development: Success for All and Reading Recovery.

Success for All is a school-level elementary school intervention that focuses on improving literacy outcomes for all students to improve overall student achievement and is currently used in 1200 schools across the country (Borman et al., 2007). The program is designed to identify and address deficiencies in reading skills at a young age using a variety of specific instruction strategies, ranging from cooperative learning to data-driven instruction. Success for All is purchased as a comprehensive package, which includes materials, training, ongoing PD, and a well-specified “blueprint” for delivering and

sustaining the model. Schools that elect to adopt Success for All implement a program that organizes resources to attempt to ensure that every child will reach the third grade on time with adequate basic skills and will continue to build on those skills throughout the later elementary grades.

Borman et al. (2007) use a cluster randomized trial design to evaluate the impacts of the Success for All model on student achievement. Forty-one schools from 11 states volunteered and were randomly assigned to either the treatment or control groups. Treatment schools implemented Success for All in grades K-2 and control schools implemented the program in grades 3–5. Borman et al. (2007) present results 3 years after randomization for the baseline cohort of kindergarten students. Although 41 schools were initially randomized, only 35 schools were included in the analysis due to six schools dropping out over the years for various reasons. The authors conclude that these 35 schools are still balanced and attrition is not a threat to their results. Using standardized test scores from three subtests of the Woodcock Reading Mastery Test—Revised, Borman et al. (2007) find that Success for All increased student achievement by  $0.36\sigma$  (0.11) on phonemic awareness,  $0.24\sigma$  (0.11) on word identification, and  $0.21\sigma$  (0.09) on passage comprehension.

Another similar professional development program is Reading Recovery (RR). RR is a short-term intervention designed to help struggling readers in first grade catch up to their peers. The program consists of students meeting one-on-one with a specially trained teacher everyday for a 30-min lesson over 12–20 weeks. The lessons are individualized by the RR teacher to fit to a student’s strengths and needs and follow the RR model—focusing on phonemic awareness, phonics, vocabulary, fluency, and comprehension. RR teachers undergo a year-long training procedure that takes place at designated training facilities and the schools where they are assigned. Through this training, they learn how to design and deliver daily lesson plans, document lessons, and to collect and effectively use different types of student progress data. All RR teachers are overseen by a teacher leader who has attended an intensive postgraduate program where they are expected to emerge as literary experts. Literature on RR reports that approximately 75% of students enrolled in RR typically reach grade-level proficiency after participating in RR for the program’s intended length of 12–20 weeks and that these students go on to maintain their progress through the remainder of elementary school (May et al., 2013).

Schwartz (2005) conducted the first randomized evaluation of RR in the United States. In his study, 37 first-grade teachers from across the nation identified two at-risk students in their classroom. One student from each pair was randomly assigned to a treatment condition that received RR in the fall and the other student was assigned to a control condition that received RR in the spring. The participating teachers were all certified RR teachers and the program was active in their schools. The teachers gave up one of their four 30-min RR slots to whichever student was randomly assigned to treatment.

At the end of the first semester, [Schwartz \(2005\)](#) found that treatment students had large and significant impacts on various observation survey and standardized reading measures. Effect sizes on the text level, letter identification, concepts about print, and hearing and recording sounds in words tasks on the observation survey ranged from  $0.9\sigma$  to  $2.02\sigma$  and treatment had an impact of  $0.94\sigma$  on scores from the Slosson Oral Reading Test—Revised.

In 2010, the US Department of Education awarded RR a \$45 million i3 grant along with \$10.1 million from private sources to fund a scale-up of RR across the nation. The scale-up intends to reach over 2000 schools and provide literary assistance to over 88,000 students. [May et al. \(2013\)](#) report the findings from the first 2 years of this scale-up.

A total of 628 schools from across the nation were enrolled in the i3 scale-up of Reading Recovery. These schools were randomly assigned to three blocks and one of these blocks was randomly chosen to participate in a randomized controlled trial (RCT) of Reading Recovery during the 2011–2012 school year. Of the 209 schools in this block, only 156 schools actually carried out the randomization process described below and were included in the evaluation. Each school that participated in the RCT identified the eight lowest scoring students in their school using the Observation Survey of Early Literacy. These eight students were matched according to their scores and ELL status and then one student from each pair was randomly assigned to treatment and the other to control. This process resulted in 628 students in the treatment group and 625 students in the control group (when there were less than eight eligible students, odd number students were automatically assigned to treatment. These students were omitted from the impact analysis).

Using standardized test scores from the Iowa Test of Basic Skills and baseline demographics, [May et al. \(2013\)](#) investigated the causal impact of being assigned to the RR program. They find that RR increased student achievement by  $0.60\sigma$  on the reading words subtest and  $0.61\sigma$  on the reading comprehension subtest. The effects of these “managed” PD experiments for both subjects are statistically significant and for reading, quite large. The metacoefficient is  $0.052\sigma$  (0.016) for math achievement and  $0.403\sigma$  (0.120) for reading achievement.

**3.3.2.3 Teacher feedback** The modernization of teacher evaluation systems, an increasingly common component of teacher professional development, promises to reveal new, systematic information about the performance of individual classroom teachers. Yet while states and districts race to design new systems, most discussion of how the information might be used has focused on traditional human resource—management tasks, namely, hiring, firing, and compensation. By contrast, very little is known about how the availability of new information, or the experience of being evaluated, might change teacher effort and effectiveness. [Fryer and Dobbie \(2013\)](#) report that teacher feedback is one of the variables most correlated with charter school success.

In the research reported here, we study one approach to teacher feedback: practice-based assessment that relies on multiple, highly structured classroom observations conducted by experienced peer teachers and administrators. While this approach contrasts with principal walk-through styles of class observation, its use is on the rise in new and proposed evaluation systems in which rigorous classroom observation is often combined with other measures, such as teacher value-added based on student test scores.

Proponents of evaluation systems that include high-quality classroom observations point to their potential value for improving instruction (see “Capturing the Dimensions of Effective Teaching,” *Features*, Fall 2012). Individualized, specific information about performance is especially scarce in the teaching profession, suggesting that a lack of information on how to improve could be a substantial barrier to individual improvement among teachers. Well-designed evaluations might fill that knowledge gap in several ways. First, teachers could gain information through the formal scoring and feedback routines of an evaluation program. Second, evaluation could encourage teachers to be generally more self-reflective, regardless of the evaluative criteria. Third, the evaluation process could create more opportunities for conversations with other teachers and administrators about effective practices.

[Taylor and Tyler \(2012\)](#), using a quasiexperimental design, find that teachers are more effective at raising student achievement during the school year when they are being evaluated as opposed to previous years, and even more effective in the years after evaluation. A student instructed by a teacher after that teacher has been through an evaluation scored about 11% of a standard deviation (4.5 percentile points for a median student) higher in math than a similar student taught by the same teacher before the teacher was evaluated.

### **3.3.3 School Management**

[Bloom et al. \(2015\)](#) identify an interesting relationship between management quality of 1800 high schools from eight countries and student achievement in those schools. They find a strong correlation between higher management quality and better educational outcomes. [Fryer and Dobbie \(2013\)](#) use variation in the management practices of New York City charter schools to investigate the characteristics that differentiate those that increase student achievement (as measured by standardized test scores) and those that do not increase achievement. Using survey and administrative data from 39 New York City charter schools, they correlated the policies of each school with the school’s individual impact on math and reading achievement. [Fryer and Dobbie \(2013\)](#) report that traditional inputs (i.e., class size, per pupil expenditure, the fraction of teachers with no certification, and the fraction of teachers with an advanced degree) are not correlated with school effectiveness. Instead, they found that frequent teacher feedback, the use of data to guide instruction, high-dosage tutoring, increased instructional time, and high expectations are highly correlated with schools’ impacts on math and reading. In this section, we present studies

that explore causal impacts of some of the management practices discussed in [Bloom et al. \(2015\)](#) and [Fryer and Dobbie \(2013\)](#).

### 3.3.3.1 Using data to drive instruction

[Carlson et al. \(2011\)](#) present results from a large randomized study that investigates the impacts of data-driven reform on student achievement in mathematics and reading. The study included over 500 schools from 59 school districts across seven states. Districts randomly assigned to treatment implemented a 3-year data-driven reform initiative with the support of consultants from the Johns Hopkins Center for Data-Driven Reform in Education (CDDRE). Control districts implemented the same initiative, but one year after random assignment. [Carlson et al. \(2011\)](#) utilize the delayed start to investigate the causal impacts of the first year of the CDDRE initiative on student achievement outcomes. The first year of the CDDRE initiative focuses on developing and evaluating quarterly benchmark assessments, reviewing all available data to better understand the needs of the district, and conducting leadership and data interpretation training for district and school leaders.

The participating districts were selected through an extensive recruitment process. The Department of Education of each state nominated districts with a large number of low performing schools to participate in the study. District officials of the nominated districts were contacted and those that agreed to participate were included in the randomization procedure. Further, for each participating district, the district officials specified which schools in their district they wanted to participate in the experiment. Generally, low performing schools were selected. Following the selection of schools, districts were stratified by recruitment wave and state and randomly assigned to treatment or control. Treatment schools implemented CDDRE data-driven initiative and control schools continued business as usual for one year and then implemented the same initiative.

To assess the impact of the intervention, results from state-administered achievement tests were collected for each participating school. [Carlson et al. \(2011\)](#) found treatment had a significant impact on student math scores but found no significant effect for reading scores — treatment schools increased students' math scores by  $0.059\sigma$  (0.029) and increased students' reading scores by  $0.033\sigma$  (0.020).

In addition to providing constructive feedback for teachers, collecting teacher data could also be a useful tool for leaders in managing their schools. [Rockoff et al. \(2012\)](#) investigate the impact of giving over 200 New York City principals objective performance evaluations of the teachers in their schools. All schools in NYC that contained any grades 4 through 8 were eligible to participate (over 1000 schools); of the above, 223 signed up and completed the necessary survey to

be included in the experiment. Participating principals were stratified by grade configuration and assigned randomly to treatment or control. Treatment principals received reports detailing the value-added of the teachers in their school relative to similar teachers in NYC and training on how to use and interpret this data. Rockoff et al. (2012) find evidence that principals do use this information to update their beliefs of the teachers in the school. Using baseline and postintervention surveys that solicited principals' evaluation of their teachers, they find that treatment principals update their beliefs in the direction of the teacher value-added detailed in the report. Moreover, consistent with a Bayesian learning model, principals put more weight on the teacher value-added information when that information is more precise than their prior beliefs and they put more weight on their prior beliefs when the relative precision is reversed. Providing this information to principals led to an increase in turnover for teachers with low performance estimates and had a positive impact on students' math achievement for students assigned to teachers that remained in the intervention throughout its entirety.

### 3.3.3.2 Class size

Project STAR was an experiment carried out in 79 Tennessee schools from 1985 to 1989 where 11,600 students in grades K to 3 were randomly assigned to small classes (13–17 students), regular classes (22–25 students), or regular classes with a full-time aide. At the time, the statewide pupil-to-teacher ratio was 22.3, so regular classes represented close to the average classroom size in the state. At the time of the experiment, kindergarten was not compulsory in Tennessee, so many new students entered schools in first grade. Students who entered a participating school after the 1985–1986 school year were randomly assigned to one of the three types of classes. Additionally, students in regular classes and in regular classes with an aide were randomly reassigned between these two types of classes at the end of kindergarten. However, kindergartners initially assigned to small classes remained in small classes throughout the entire experiment.

Using a student's initial assignment to one of the three groups, Krueger (1999) estimated the impact of reduced class size and teacher aides on an index of scores from the math, reading, and word subtests of the Stanford Achievement Test. Krueger (1999) found that for grades K–3, students scored about 5–7 percentile points higher on the index than students assigned to a regular class without an aide. These results correspond to effect sizes in the range of  $0.19\sigma$ – $0.28\sigma$  and represent 64–82% of the white–black test score gap in the data. Additionally, there was some evidence that regular classrooms with aides outperformed regular classrooms without aides—the estimates for aide classrooms tended to be small and positive, but only the first grade results were statistically significant

with an impact of 1.48 percentile points. When exploring heterogeneous treatment effects, Krueger (1999) found that smaller class sizes were more effective for students on free lunch and black students.

### 3.3.3.3 Extended time

There are very few randomized trials that expose students to higher quantities of schooling. Zvoch and Stevens (2012) show that a summer literacy program has enormous impacts on kindergarten and first grade reading test scores. In this study, the researchers invited students to a 5-week summer program that lasted for 3.5 h a day, 4 days a week. In the program, students received classroom instruction on fundamental literacy topics, were assigned homework, completed in-class work packets, and practiced literacy skills in small groups with students of a similar skill level. The summer program was typically reserved for struggling students that scored below a cutoff point on the spring standardized tests. However, for this study, the district established upper bounds so that approximately 50 kindergartners and 50 first graders fell in the range between the cutoff scores and the upper bound scores. These students were considered the experimental sample and half were randomly invited to participate in the program. At posttest, Zvoch and Stevens (2012) found that the summer program on average increased reading test scores by  $0.69\sigma$  for the kindergarten and first grade students.

However, Holmes and McConnell (1990) utilized a larger sample of students to investigate the impact of full-day versus half-day kindergarten instruction and found no significantly positive impacts. In fact, their study provided evidence that half-day kindergarten students perform better on math achievement tests than full-day kindergartners. The experiment randomly assigned 20 elementary schools to either a full or half-day schedule. Holmes and McConnell found that full-day kindergartners had math scores that were  $0.29\sigma$  lower and reading scores that were  $0.11\sigma$  higher than the half-day students.

An experiment that investigated extended day impacts in a slightly older sample was Meyer and Van Klaveren (2013). This experiment randomly invited Dutch fifth-, sixth-, and seventh-grade students to participate in an extended school day program. The program consisted of a classroom of approximately 10 students receiving an additional 2 h of language instruction, 2 h of math instruction, and 1 h of excursions per week. Meyer and Van Klaveren found that assignment to treatment increased math scores by  $0.087\sigma$  ( $0.067$ ) and increased reading scores by  $0.005\sigma$  ( $0.081$ ). Neither of these effects is significant.

Taking the treatment effects at face value, one potential explanation for the patterns in the experimental data is that increasing the amount of time students spend in class per day is not as effective as extending the school year. Put differently, if there are concavities in human capital production as a function of time and students are at the point of diminishing marginal returns for a given day but not for a given year, this can rationalize the findings.

### 3.3.4 Market-based approaches

In recent years, developed countries across the globe have increased the scope of schooling alternatives available to students—an approach long advocated by leading economists (Friedman, 1955; Becker, 1995; Hoxby, 2002). Creating a competitive and active marketplace has the potential to improve educational outcomes because schools would have more incentive to improve in response to increased market pressure. To the extent that match quality between a school and a student is important, school choice programs may also yield benefits simply by increasing the set of schools over which a student is able to choose.

For these approaches to be an effective means of reform, however, it is necessary that students benefit from the opportunity to attend sought-after schools, and that these improvements are apparent to students and parents. The goal of this subsection is to understand the measurable achievement benefits accrued to students when there is more flexibility and school choice.

#### 3.3.4.1 Vouchers

There have been a series of important studies that exploit randomized voucher lotteries to estimate the effect of attending a private school for youth at various ages. The Milwaukee voucher program, offering vouchers to a limited number of low-income students to attend one of three private nonsectarian schools in the district, is the most prominent of these. Analyses of this program obtain sharply conflicting estimates of the impact on achievement depending upon the assumptions made to deal with selective attrition of lottery losers from the sample (Witte et al., 1995; Greene et al., 1999; Witte, 1997; Rouse, 1998). Although in theory randomization provides an ideal context for evaluating the benefits of expanding parental choice sets, in the Milwaukee case, less than half of the unsuccessful applicants returned to the public schools and those who did return were from less educated, lower income families (Witte, 1997).

Rouse (1998) used a typical ITT specification to evaluate the Milwaukee voucher program. Comparing lottery winners to lottery losers, she found that being selected for the choice program had significant impacts on math achievement but insignificant impacts on reading. Students who won the lottery scored approximately 1.5–2.3 percentile points ( $0.08\sigma$ – $0.12\sigma$ ) more per year in math compared to lottery losers. This suggests effect sizes on the order of  $0.32\sigma$ – $0.48\sigma$  for 4 years of school. The results in Rouse (1998) are robust to various methods of imputing missing data and attrition from the sample—when imputing missing observations, estimates remained in the range of 1.38–2.31 percentile points.

The DC Opportunity Scholarship Program (OSP) is another voucher program that provides up to \$7500 to low-income families in the District of Columbia to send their children to participating private schools. Wolf et al. (2010) use 2300 applicants to a

series of lotteries in 2004 and 2005 to evaluate the impact of the OSP. The study found that the OSP had no impact on student achievement but increased students' chance of graduation. Additionally, parents of students who were offered a scholarship had a higher satisfaction with schools and rated schools as safer. This result is significant regardless of whether a student actually used the offered scholarship or not.

[Mayer et al. \(2002\)](#) present results from the third year of a randomized evaluation of the School Choice Scholarships Foundation Program in NYC. In 1997, the program provided scholarships of up to \$1400 annually for up to four years via lottery to low-income families with students in grades K–4. The scholarship could be used to pay tuition at a religious or secular school of the family's choosing. Fifty-three percent of students who were offered scholarships used the scholarship for at least 3 full years. The families that did not utilize the offered scholarship claimed they were unable to do so because they were unable to afford the tuition and expenses that the scholarship did not cover or were unable to find a school in a convenient location.

Through parent and student surveys, [Mayer et al. \(2002\)](#) found that the private schools these students elected to attend were indeed different from the public schools nonparticipants remained in. Parents with students who switched to private schools reported that the schools had smaller class sizes; were more likely to have computer laboratories, after-school programs, and tutor programs; had fewer incidents of students destroying property, fighting, cheating, and racial conflict; communicated more with parents; allowed parents to spend less time helping their children with homework; and this resulted in an overall higher level of satisfaction with their students' school. Students who switched reported that students in private school were more likely to get along with teachers, were more proud of their school, were less likely to be put down by teachers, and were asked to complete more homework. Additionally, students reported that the private schools had stricter behavior rules, and there was a lower prevalence of cheating.

Although there was evidence that students offered a scholarship switched to better school environments, [Mayer et al. \(2002\)](#) found that three years after random assignment, there was no average treatment effect on students' performance. Moreover, students who ever attended a private school and students who attended for all 3 years did no better than students who never attended a private school. These results are robust across grade levels, but there is evidence of heterogeneous treatment effects across races — [Mayer et al. \(2002\)](#) found positive effects on the standardized test scores of black students.

The metacoefficient on voucher experiments is  $0.024\sigma$  (0.021) for math achievement and  $0.030\sigma$  (0.024) for reading achievement. Relative to their popularity with politicians, the lack of effectiveness of voucher programs is surprising. Rather than focusing on achievement, many use a revealed preference argument to conclude families who make active choices — even if achievement is unaffected — are better off.

Before one dismisses them entirely, there are two key pieces of data missing on voucher experiments. First, in the average voucher experiment a student enrolls in a private school between grades K and 8. There is no experiment that tests the full pre-K through high school graduation treatment. This seems essential.

Second, although vouchers are a market-based reform, we do not know what happens if there are enough vouchers in a concentrated area to allow the market to respond by altering the supply (and scope) of schools available to educate disadvantaged children. Because all the experiments have been relatively small, one cannot assess the potential general equilibrium effects.<sup>28</sup>

An ideal voucher experiment might take a large state with multiple school districts and randomly implement voucher programs or Education Savings Accounts in half of the districts and analyze both student achievement and the market response. The vouchers could be risk adjusted — more disadvantaged children receive more school funding — or contain location preferences that would induce a more aggressive supply response in blighted communities. These ideas only scratch the surface of what is possible and have not been evaluated in a compelling way. Thus, whether the [Friedman \(1955\)](#) vision for public schools is effective at producing human capital is still unknown.

### 3.3.4.2 School choice

[Cullen et al. \(2006\)](#) present causal estimates of the impact of school choice on a variety of student outcomes. Specifically, they utilize the random lotteries of oversubscribed schools in Chicago's open enrollment system. This system allows students to apply to public magnet schools and programs outside of their neighborhood school.<sup>29</sup> When oversubscribed, many Chicago Public Schools (CPS) use random lotteries to offer admission to students. The authors obtained the results of 194 such lotteries from 19 high schools in CPS. The final sample consisted of 14,434 students who applied to these 19 choice schools in the spring of 2000 and 2001.

<sup>28</sup> A notable counter example is a recent experiment implemented in India. [Muralidharan and Sundararaman \(2011\)](#) conducted an experiment using 180 villages from the Indian state of Andhra Pradesh in which they randomly assigned villages to treatment or control and then awarded private school vouchers to public school applicants through random lotteries in the treatment villages. Two and four years after randomization, they found that winning a voucher had no impact on Telugu (native language), math, English, and science/social studies achievement. However, the program had large impacts on Hindi test scores, a subject not taught in public schools. Since private schools are approximately a third of the cost of public schools, [Muralidharan and Sundararaman \(2011\)](#) conclude that private schools are a much more cost-effective way of teaching students. Further, they found no evidence of spillovers (negative or positive) on the achievement of public school students that did not apply to the voucher program or on nonvoucher private students. This suggests that vouchers are a cost-effective way to potentially increase student achievement without any negative externalities.

<sup>29</sup> Magnet schools are different from traditional public schools in that each magnet school tends to have a specific educational theme and students can choose to enroll in a school based on their interest in a school's theme.

The analysis in [Cullen et al. \(2006\)](#) finds little evidence that winning a lottery has any impact on traditional achievement measures such as test scores, graduation rates, attendance rates, or courses taken. These results are robust to a variety of sensitivity analyses and are similar across student subgroups. In an attempt to better understand the findings, the authors explored potential mechanisms that could explain the zero-impact on academic outcomes. They found little evidence of lottery winners and losers attending similar schools (lottery winners attended schools with higher average achievement, lower poverty rates, and higher graduation rates), of choice schools substituting for parental involvement, or of travel costs and disruption of peer groups interfering with academic success. Therefore, the results in [Cullen et al. \(2006\)](#) seem to suggest that the measurable school inputs of these choice schools have little causal impact on students' academic outcomes.

Another possibility is that students and parents apply to choice schools for nonacademic reasons. Using survey data collected by the Consortium on Chicago School Research for CPS students in grades 6–10 in spring 2001, the authors investigated this possibility. They found evidence of some positive effects on nontraditional outcomes, possibly supporting the hypothesis that students and parents choose choice schools for nonacademic reasons. [Cullen, Jacob, and Levitt \(2006\)](#) found that lottery winners report fewer incidents of disciplinary action, fewer arrests, and lower incarceration rates. However, lottery winners are not statistically different from lottery losers for other outcomes such as liking school, trusting their teachers, and having high expectations for the future.

Another example of a school choice experiment is Connecticut's interdistrict magnet school program. In 1996, the Connecticut Supreme Court ruled that students in Hartford public schools were denied equal educational opportunities due to racial and economic isolation. One of the state's many responses was to foster the growth of interdistrict magnet schools. A decade after the Connecticut Supreme Court's ruling, there were 54 magnet schools in operation in Connecticut and 41 of these served students residing in Hartford, New Haven, or Waterbury. Additionally, interdistrict magnets serve two or more districts and all students residing in these districts are eligible to enroll in the school. Urban students that elect to attend magnet schools are typically moving to schools where there are fewer students eligible for free lunch, more white students, and higher average scores on standardized mathematics and reading tests.

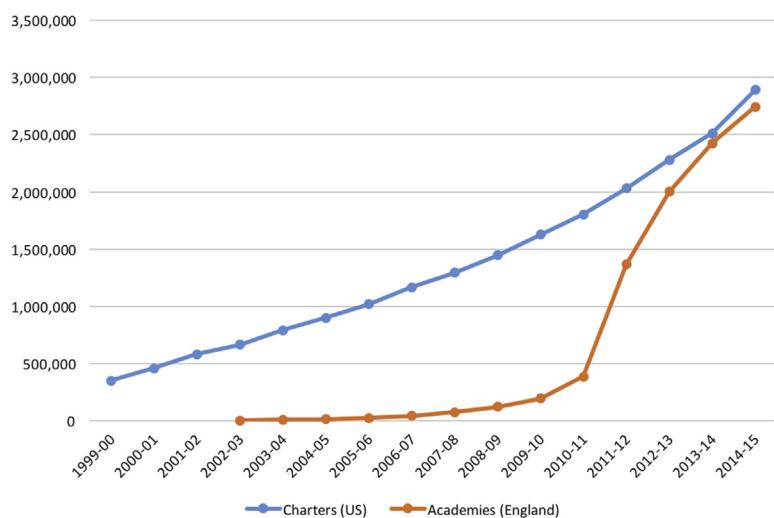
[Bifulco et al. \(2009\)](#) evaluated the impact of Connecticut's interdistrict magnet schools using the random admission lotteries of two oversubscribed magnets serving Hartford and four surrounding suburban districts. One of these schools served grades 6–8 and the other served grades 6–12. The authors obtained admission data for the 2003 and 2004 6th grade lotteries at these schools as well as student-level test scores from the Connecticut State Department of Education for the 2001–2002 to

2006–2007 school years. The final sample for these two schools consisted of 553 students in 12 oversubscribed lotteries (both schools conducted lotteries by district for each year), 164 of which were eventually offered admission to one of the two magnets. Comparing the 8th grade outcomes of lottery winners to lottery losers, [Bifulco et al. \(2009\)](#) find that students offered admission to the magnet schools scored  $0.109\sigma$  higher on math and  $0.252\sigma$  higher on reading tests, of which only the latter was statistically significant at conventional levels.

### 3.3.4.3 Charter schools

A charter school is a school that receives public funding but operates independently of the established public school system in which it is located. They exist (and are increasing in demand) across the developed world — from Australia to England and Wales. [Fig. 3](#) shows the increase in the number of students attending charter schools in the United States and England.

When originally conceived, charter schools offered two distinct promises: First, they were to serve as an escape hatch for students in failing schools. Second, they were to use their legal and financial freedoms to create and incubate new



**Figure 3 Number of students in “charters.”** This figure presents the number of students enrolled in charter schools (US) and academies (England) for the 1999–2000 to 2014–15 school years. Academies are publicly funded independent schools. Similar to the US charter schools, academies do not have to follow the national curriculum and term lengths. The US data comes from the National Alliance for Public Charter Schools (NAPCS) and the data for England comes from the UK Department of Education. Note that the US 2014–2015 number is estimated ([NAPCS, 2015](#)).

educational practices that could be used to inform traditional public schools with new ideas and fresh approaches.

In America, charter schools currently enroll almost 4% of all students. Some of these schools have shown remarkable success in increasing test scores — closing the racial achievement gap in just a few years. For example, schools such as the Success Academy Charter Schools in New York City, YES Prep in Houston, and charter schools in the Harlem Children’s Zone have become beacons of hope, demonstrating the enormous potential to improve student achievement in the most blighted communities. Others, however, have failed to increase achievement and have actually performed substantially worse than their traditional counterparts. In this scenario, students would have been better off not attending a charter school.

**3.3.4.3.1 Evaluating charter schools** The method for evaluating charter schools is remarkably consistent across the literature.<sup>30</sup> The literature estimates two empirical models — ITT effects and Local Average Treatment Effects (LATEs)—which provide a set of causal estimates of the impact of attending a charter school on outcomes. The ITT estimates measure the causal effect of winning a charter lottery by comparing the average outcomes of students who “won” the lottery to the average outcomes of students who “lost” the lottery:

$$\text{outcome}_i = \mu + \gamma X_i + \pi Z_i + \sum_j \nu_j \text{Lottery}_{ij} + \sum_j \phi_j \text{Lottery}_{ij} \times 1(\text{sibling}_i) + \eta_i \quad (1)$$

where  $Z_i$  is an indicator for winning an admissions lottery, and  $X_i$  includes controls for student-level demographics such as gender, race, special education status, eligibility for free or reduced-price lunch, receipt of Limited English Proficiency services, and a quadratic in two prior years of math and ELA test scores.  $\text{Lottery}_{ij}$  is an indicator for entering the lottery in year  $j$ , and  $1(\text{sibling}_i)$  indicates whether student  $i$  had a sibling enter the lottery in the same year.<sup>31</sup> Eq. (1) identifies the impact of *being offered a chance* to attend a charter school,  $\pi$ , where the lottery losers form the control group corresponding to the counterfactual state that would have occurred for students in the treatment group if they had not been offered a spot in the charter

<sup>30</sup> The national charter school studies released by the Center for Research on Education Outcomes (CREDO) are anomalous in that they use observational data instead of randomized admissions lotteries ([Center for Research on Education Outcomes, 2013](#)).

<sup>31</sup> In typical charter lotteries, an offer is extended to all siblings when multiple siblings enter the same lottery and one sibling wins.

school. Using this approach, the literature on charter effectiveness has quickly amassed an interesting set of facts.

First, the typical charter school is no more effective at increasing test scores than the typical traditional public school (Gleason et al., 2010). Evaluations that encompass the most representative samples of charter schools show little impact. Using lottery admissions data for 36 charter schools from around the nation, Gleason et al. (2010) investigated the impact of charter schools on student outcomes. They found that 2 years after the random lotteries, students who won the lotteries scored, if anything, lower on standardized test scores than students who lost the lotteries.<sup>32</sup> In addition, this national sample of charter schools had no impact on students' math and reading proficiency levels, number of days absent, and grade promotion. Gleason et al. (2010) found no impact of charter schools on student behavior and school disciplinary action, but a higher fraction of lottery winners showed up late to school five or more days. Although the average charter school included in their study did not have any positive impacts on student outcomes, Gleason et al. (2010) found large, positive, and statistically significant impacts — ranging from  $0.07\sigma$  to  $0.94\sigma$  — on every measure of students' and parents' satisfaction with and perceptions of school.

Second, an emerging body of research suggests that high-performing charter schools can significantly increase the achievement of poor urban students. Students attending over-subscribed Boston-area charter schools score approximately  $0.4\sigma$  higher per year in math and  $0.2\sigma$  higher per year in reading (Abdulkadiroglu et al., 2011). Promise Academy students in the Harlem Children's Zone score  $0.229\sigma$  higher per year in math and  $0.047\sigma$  higher per year in reading (Dobbie and Fryer, 2011). Students in the Knowledge is Power Program (KIPP) schools — America's largest network of charter schools — score  $0.180\sigma$  higher per year in math and  $0.075\sigma$  higher per year in reading (Tuttle et al., 2013; Angrist et al., 2011). The SEED urban boarding school in Washington DC demonstrates similar test score gains (Curto and Fryer, 2014).

Third, charter schools are more effective at increasing math scores than reading scores. Abdulkadiroglu et al. (2011) and Angrist et al. (2011) find that the treatment effect of attending an oversubscribed charter school is four times as large for math as reading. Dobbie and Fryer (2011) demonstrate effects that are almost 5 times as large in middle school and 1.6 times as large in elementary school in favor of math. In larger samples, Hoxby and Murarka (2009) report an effect size 2.5 times as large in New York City charters, and Gleason et al. (2010) show that an average urban charter school increases math scores by  $0.16\sigma$  with statistically 0 effect on reading.

<sup>32</sup> The two-year ITT impact for the pooled sample is  $-0.08\sigma$  ( $p$ -value = 0.032) for reading scores and  $-0.06\sigma$  ( $p$ -value = 0.136) for math test scores. Note that pooling results together masks heterogeneous treatment effects described in the paper. For example, charter schools in large urban areas had a  $0.16\sigma$  impact on math scores while schools outside of large urban areas had a  $-0.14\sigma$  impact. Both of these impacts were statistically significant.

According to the [National Alliance for Public Charter Schools \(2009\)](#), the median grade served by charter schools in the United States is 6th grade (usually students are 11–12 years old). However, the achievement data necessary to conduct evaluations of charter schools is typically not available for kindergarten through 2nd grade students, so the average grade evaluated is most likely even higher. The theory and empirical findings discussed above suggest that the relatively late timing of charter school “interventions” might be an important factor in the observed differential impacts by subject.

Another leading theory posits that reading scores are influenced by the language spoken when students are outside of the classroom ([Rickford, 1999](#); [Charity et al., 2004](#)). [Charity, Scarborough, and Griffin \(2004\)](#) argue that if students speak nonstandard English at home and in their communities, increasing reading scores might be especially difficult. This theory is consistent with the data and could explain why students at an urban boarding school make similar progress on reading and math ([Curto and Fryer, 2014](#)).

Fourth, there are important features of charter schools that seem to be correlated with their level of student achievement. It is important to note that these analyses are nonexperimental. [Angrist et al. \(2013\)](#) argue that both the urbanicity of charter schools and whether they adopt the so-called “No Excuses” approach to culture and discipline are positive predictors of charter treatment effects.

[Fryer and Dobbie \(2013\)](#) provide evidence on the determinants of charter school effectiveness by collecting data on the inner workings of 29 charter schools in New York City and correlating these data with lottery-based estimates of each school’s effectiveness. Information on school practices was collected from a variety of sources. Principal interviews asked about teacher development, instructional time, data driven instruction, parent outreach, and school culture. Teacher interviews asked about professional development, school policies, school culture, and student assessment. Student interviews asked about school environment, school disciplinary policy, and future aspirations. Lesson plans were used to measure curricular rigor. Videotaped classroom observations were used to calculate the fraction of students on task throughout the school day.

School effectiveness is estimated by exploiting the fact that oversubscribed charter schools in New York City are required to admit students via random lottery. The variability inherent in the set of NYC charter schools, combined with rich measures of school inputs and lottery-based estimates of each school’s impact on student achievement, provides an ideal opportunity to understand which inputs best

explain school effectiveness. This, coupled with some of the best practices of our meta-analysis, provide the intellectual backbone of the randomized field trial discussed below.

Fryer and Dobbie (2013) find that input measures associated with a traditional resource-based model of education — class size, per pupil expenditure, the fraction of teachers with no teaching certification, and the fraction of teachers with an advanced degree — are not correlated with school effectiveness in our sample. Indeed, our data suggest that increasing resource-based inputs may marginally lower school effectiveness. On the surface, this evidence may seem inconsistent with the important results reported in Krueger (1999). There are a few ways to reconcile this. First, Fryer and Dobbie (2013) analyzes charter schools in NYC, whereas Krueger (1999) uses a sample of traditional public schools in Tennessee. Second, the variation in Fryer and Dobbie (2013) comes from only 39 charter schools with relatively similar class sizes, whereas the thousands of treatment students in Krueger (1999) are placed in classrooms that are almost 40% smaller than control classrooms. Third, Krueger's analysis focused on students in grades kindergarten through third whereas the correlations in Fryer and Dobbie (2013) used third through 8th grade test scores. Fourth, and most important, the analysis in Krueger (1999) is experimental.

In stark contrast, Fryer and Dobbie (2013) demonstrate that an index of five policies suggested by 40 years of human capital research—frequent teacher feedback, data-driven instruction, high-dosage tutoring, increased instructional time, and a relentless focus on academic achievement—explains roughly half of the variation in school effectiveness in both math and reading.

#### **4. COMBINING WHAT WORKS: EVIDENCE FROM A RANDOMIZED FIELD EXPERIMENT IN HOUSTON**

Improving the efficiency of the production of human capital is of great importance across the developed world. The United States spends \$10,768 per pupil on primary and secondary education, ranking it fourth among OECD countries (Aud et al., 2011). Yet, among these same countries, American 15 year-olds rank twenty-fifth in math achievement, seventeenth in science, and fourteenth in reading (Fleischman, 2010). This is not a phenomenon that is unique to the United States. Other OECD countries are unable to translate large amounts of educational spending into educational success. For example, the two countries ranking directly behind the United States with per pupil primary and secondary spending of \$9959 and \$9448 are, respectively, Austria and

Denmark ([Aud et al., 2011](#)). However, Austrian 15 year-olds rank eighteenth in math achievement, twenty-fourth in science, and thirty-first in reading and Danish 15 year-olds rank thirteenth in math, twentieth in science, and nineteenth in reading ([Fleischman, 2010](#)).

Traditionally, there have been two approaches to increasing educational efficiency: (1) expand the scope of available educational options in the hope that the market will drive out ineffective schools, or (2) directly manipulate inputs to the educational production function.<sup>[33](#)</sup>

As our meta-analysis demonstrates, market-based reforms such as school choice or school vouchers have, at best, a modest impact on student achievement. This suggests that these approaches—implemented in their current form—are unlikely to significantly increase the efficiency of the public school system, subject to the important caveats discussed in the previous section.

Another approach is to inject the best practices known from the set of randomized field experiments completed to date—along with the correlates gleaned from analyzing the inner-workings of successful charter schools—in an experiment in traditional public schools. This is precisely the goal of [Fryer \(2014a\)](#).

Between the 2010–2011 and 2012–2013 school years, [Fryer \(2014a\)](#) implemented important elements of the above education best practices in 20 of the lowest performing schools (containing more than 12,000 students) in Houston, Texas.

To increase time on task, the school day was lengthened by 1 h and the school year was lengthened by 10 days in the nine secondary (middle and high) schools. This was 21% more time in school than students in these schools spent in the pretreatment year and roughly the same as achievement-increasing charter schools in New York City. In addition, students were strongly encouraged and even incentivized to attend classes on Saturday. In the 11 elementary schools, the length of the day and the year were not changed, but noninstructional activities (e.g., 20-min bathroom breaks) were reduced. This is consistent with the correlations in [Fryer and Dobbie \(2013\)](#) and the randomized field trial reported in [Meyer and Van Klaveren \(2013\)](#).

In an effort to improve the human capital available to teach students and lead schools, 19 out of 20 principals were removed and 46% of teachers left or were removed before the experiment began. Some teachers left because they believed the program was too disruptive. Others were removed because they were too resistant to the changes. Any

<sup>33</sup> Increasing standards and accountability reflect a third approach to education reform. There is evidence that increased accountability via the No Child Left Behind Act had a positive impact on math test scores (though not reading test scores) and on wages ([Dee and Jacob, 2011; Deming et al., 2013](#)).

teacher, independent of skill level, who demonstrated a desire to implement the proposed changes with fidelity, was retained. As part of the turnaround efforts, teachers received both managed professional development and frequent feedback as a part of a more holistic evaluation system. The managed professional development was similar to the Success for All treatment described in [Borman et al. \(2007\)](#). The frequent feedback was similar to the quasiexperimental program evaluated in [Taylor and Tyler \(2012\)](#).

To enhance student-level differentiation, all 4th, 6th and 9th graders received high-dosage math tutoring and extra reading or math instruction was provided to students in other grades who had previously performed below grade level. Similar to the Chicago BAM experiment described above, the tutoring model was adapted from the MATCH school in Boston — a charter school that largely adheres to the methods described in [Fryer and Dobbie \(2013\)](#).

To help teachers use interim data on student performance to guide and inform instructional practice, schools were required to administer interim assessments every 3 to 4 weeks and provided with three cumulative benchmark assessments, as well as assistance in analyzing and presenting student performance data on these assessments. Yet, as [Rockoff et al. \(2012\)](#) and [Fryer and Dobbie \(2013\)](#) demonstrate, data alone is not enough. [Fryer and Dobbie \(2013\)](#) argue that the use of interim assessment data is only correlated with achievement for schools who can articulate a precise plan of how they will change student grouping or pedagogy or some other strategy in response to the data.

Finally, to instill a culture of high expectations and college access, we started by setting clear expectations for school leadership. Schools were provided with a rubric for the school and classroom environment and were expected to implement school-parent-student contracts. Specific student performance goals were set for each school and the principal was held accountable and provided with financial incentives based on these goals.

Such invasive changes were possible, in part, because 11 of the 20 schools (nine secondary and two elementary) were either “chronically low performing” or on the verge of being labeled as such and subject to takeover by the state of Texas. Thus, despite our best efforts, random assignment was not a feasible option for these schools. To round out our sample of 20 schools and provide a way to choose between alternative quasiexperimental specifications, we randomly selected nine additional elementary schools (vis-à-vis matched-pairs) from 18 low—but not chronically low—performing schools. One of the randomly selected treatment elementary schools closed before the start of the experiment so we had to drop it and its matched pair from our experimental sample. Thus, our final experimental sample consists of 16 schools.

In the sample of 16 elementary schools in which treatment and control were chosen by random assignment, providing estimates of the impact of injecting charter school best practices in traditional public schools is straightforward. In the remaining set of schools, we use three separate statistical approaches to understand the impact of the intervention. Treatment is defined as being zoned to attend a treatment school for entering grade levels (e.g., 6th and 9th) or having attended a treatment school in the pretreatment year for returning grade levels. “Comparison school” attendees are all other students in Houston. We begin by using district administrative data on student demographics and, most importantly, previous years’ achievement, to fit least squares models. We then present two empirical models that instrument for a student’s attendance in a treatment school with original treatment assignment.

All statistical approaches lead to the same basic conclusions. Injecting best practices from charter schools into low performing traditional public schools can significantly increase student achievement in math and has marginal, if any, effect on English Language Arts (hereafter known simply as “reading”) achievement. Students in treatment elementary schools gain around  $0.184\sigma$  in math per year, relative to comparison samples. Taken at face value, this is enough to eliminate the racial achievement gap in math in Houston elementary schools in approximately three years. Students in treatment secondary schools gain  $0.146\sigma$  per year in math, decreasing the gap by one-half over the length of the demonstration project. The impacts on reading for both elementary and secondary schools are small and statistically zero.

In the grade/subject areas in which we implemented all five policies described in [Fryer and Dobbie \(2013\)](#) — fourth, 6th, and 9th grade math — the increase in student achievement is substantially larger than the increase in other grades. Relative to students who attended control schools, 4th graders in treatment schools scored  $0.331\sigma$  ( $0.104$ ) higher in math, per year. Similarly, 6th and 9th grade math scores increased  $0.608\sigma$  ( $0.093$ ), per year, relative to students in comparison schools.

#### **4.1 Simulating the potential impact of implementing best practices in education on wage inequality**

An important question is how much of the initial gaps described in the introduction to this chapter might be eliminated if state, local, and federal governments focused on the experiments proven most effective through randomized trials. Answering this question is, by definition, speculative — as it relies on extrapolations from cross-sectional relationships and assumptions on how human capital propagates through an individual’s life. Still, the exercise may be informative and we include it here as an illustrative exercise.

Data on long-term follow-ups is sparse. Perry Preschool, the Abecedarian Project, and the Moving to Opportunity experiments are notable exceptions. As described above, MTO revealed that despite having no significant impacts on children's academic outcomes, better neighborhoods had important impacts on the adulthood outcomes of children — treatment MTO children who were younger than 13 years old at randomization had 31% higher income, had higher college attendance rates, were less likely to be single parents, and lived in better neighborhoods relative to similar individuals in the control group. At posttest, the famous early childhood programs Perry Preschool and the Abecedarian Project had large impacts on children's achievement scores. At age 40, treatment students from Perry Preschool had higher high school completion rates (77% vs. 60%), were more likely to be employed (76% vs. 62%), had higher median annual earnings (\$20,800 vs. \$15,300), were more likely to own a house (37% vs. 28%), were more likely to have a savings account (76% vs. 50%), and had better crime outcomes, self-reported health, and family outcomes compared to the control group ([Schweinhart et al., 2005](#)). Similarly, at the age 30 follow-up, treatment students from the Abecedarian Project had significantly higher levels of educational attainment (13.46 years versus 12.31 years), were 17 percentage points more likely to hold a bachelor's degree (CM = 6%), were 22 percentage points more likely to work full-time (CM = 53%), and were six times less likely to receive public assistance for more than 10% of the preceding seven years than students who were assigned to control.

In the absence of more long term outcomes for the vast majority of randomized field trials, we follow the methods described in [Winship and Owen \(2013\)](#) and simulate a life-cycle model similar to the Social Genome Model (SGM).<sup>34</sup> The SGM is a useful tool to simulate how shocks in a given life-stage may carry over to later life outcomes. For example, one can simulate how much increasing reading test scores in early childhood by  $0.4\sigma$  would impact income at age 40. We can thus use this simulation — coupled with data on treatment effects from the meta-analysis — to investigate what sort of income benefits might accrue if we simply implement best practices. [Winship and Owen \(2013\)](#) provide evidence that the SGM reasonably replicates key adult impacts of the Perry Preschool experiment, the Abecedarian Project, and the Chicago Child-Parent Centers program. We find similar results.

<sup>34</sup> Due to there being no source code available — even upon request — and limited description in the SGM guide, we constructed the model using our own assumptions about the cleaning, creation, and merging of the data. This leads to a final dataset used for the simulations that is different from the one described in [Winship and Owen \(2013\)](#). However, when comparing the simulated impacts reported in published papers using SGM to estimated impacts of the simulations, they are quite similar. We provide the code and data in an online appendix.

### 4.1.1 Interpreting the literature through a simple life-cycle model

The model draws from the vast literature of human capital formation and assumes that cognitive and noncognitive skill formation varies across an individual's lifetime and is dependent on the stock of skills in previous stages of life. Specifically, [Winship and Owen \(2013\)](#) define six different life-stages: circumstances at birth (CAB), early childhood (EC), middle childhood (MC), adolescence (AD), transition to adulthood (TTA), and adulthood (AH). The empirical model uses linear structural equations to describe the dependencies between the outcomes in a given stage and all revealed outcomes from the stages preceding it. Formally, given a vector of circumstances at birth, CAB, for individual  $i$ , each outcome in the vector of early childhood outcomes, EC, is modeled as

$$\text{EC Outcome}_i = \beta_0^{\text{ec}} + \beta_{\text{cab}}^{\text{ec}} \text{CAB}_i + \varepsilon_i^{\text{ec}}.$$

Similarly, each of the MC outcomes is given by

$$\text{MC Outcome}_i = \beta_0^{\text{mc}} + \beta_{\text{cab}}^{\text{mc}} \text{CAB}_i + \beta_{\text{ec}}^{\text{mc}} \text{EC}_i + \varepsilon_i^{\text{mc}}.$$

For the adolescent life-stage we have

$$\text{AD Outcome}_i = \beta_0^{\text{ad}} + \beta_{\text{cab}}^{\text{ad}} \text{CAB}_i + \beta_{\text{ec}}^{\text{ad}} \text{EC}_i + \beta_{\text{mc}}^{\text{ad}} \text{MC}_i + \varepsilon_i^{\text{ad}}.$$

Outcomes when transitioning to adulthood would be

$$\text{TTA Outcome}_i = \beta_0^{\text{tta}} + \beta_{\text{cab}}^{\text{tta}} \text{CAB}_i + \beta_{\text{ec}}^{\text{tta}} \text{EC}_i + \beta_{\text{mc}}^{\text{tta}} \text{MC}_i + \beta_{\text{ad}}^{\text{tta}} \text{AD}_i + \varepsilon_i^{\text{tta}}.$$

And finally, adult outcomes are modeled as

$$\text{AH Outcome}_i = \beta_0^{\text{ah}} + \beta_{\text{cab}}^{\text{ah}} \text{CAB}_i + \beta_{\text{ec}}^{\text{ah}} \text{EC}_i + \beta_{\text{mc}}^{\text{ah}} \text{MC}_i + \beta_{\text{ad}}^{\text{ah}} \text{AD}_i + \beta_{\text{tta}}^{\text{ah}} \text{TTA}_i + \varepsilon_i^{\text{ah}}.$$

Where  $\beta_\psi^\lambda$  are the partial correlations of realized outcomes from the  $\psi$  life-stage ("0" represents an intercept) with the given LHS outcome in the  $\lambda$  life-stage.

With a rich enough dataset, one can obtain the correlations linking all CAB, EC, MC, AD, TTA, and AH outcomes together and investigate the indirect and direct impacts of varying one outcome on another. Importantly, we could then use the structural equations of this model to predict how a shock in earlier life-stages will propagate to outcomes in adulthood.

### 4.1.2 Simulating the Social Genome Model

Unfortunately, as discussed by [Winship and Owen \(2013\)](#), there is not yet a reliable dataset that follows an individual from birth through adult outcomes. Therefore, to conduct

simulations using the above model, we combine two well known public datasets: the National Longitudinal Survey of Youth 1979 (NLSY79) and the NLSY79 Child and Young Adult survey (CNLSY). From the CNLSY, we observe CAB, EC, MC and AD outcomes. From the NLSY79, we observe TTA and AH outcomes. See [Table 3](#) for a list of the specific variables that were used for each life-stage. The variables include a mix of cognitive skills (e.g., standardized test scores), noncognitive skills (e.g., self esteem and hyperactivity indices), and important life outcomes (e.g., teen birth, drug use, and graduation).

Using these two datasets and the equations above, we are able to estimate the coefficients for each outcome in a life-stage. However, an issue arises in linking the life-stages across these two data sources. Due to the age of respondents at first interview in the NLSY79, the data from earlier life stages is not as rich as in the CNLSY. Therefore, the NLSY79 does not contain all of the CAB, EC, MC, and AD variables that the CNLSY has. To overcome this, we define a set of linking variables, LINK, that contains all outcomes that are available in both the NLSY79 and the CNLSY. We can then estimate the following two equations in the NLSY79 dataset to obtain coefficients for each TTA and AH outcome:

$$\begin{aligned} \text{TTA Outcome}_i &= \beta_0^{\text{tta}} + \beta_{\text{link}}^{\text{tta}} \text{LINK}_i + \varepsilon_i^{\text{tta}} \\ \text{AH Outcome}_i &= \beta_0^{\text{ah}} + \beta_{\text{link}}^{\text{ah}} \text{LINK}_i + \beta_{\text{tta}}^{\text{ah}} \text{TTA}_i + \varepsilon_i^{\text{ah}}. \end{aligned}$$

Using all of the coefficients generated from these estimations and the CNLSY data, we can then build a synthetic baseline dataset of birth to age 40 outcomes for the CNLSY sample. Given the impact of an intervention at some life-stage, we can then use the same coefficients to propagate the effects of the intervention through the life-stages of these individuals. Comparing a postintervention estimation of an outcome to the baseline estimation would then provide us with an estimated impact of the intervention on the given outcome. See [Winship and Owen \(2013\)](#) and our Online Appendix B for a more in depth discussion of the estimation process.

#### **4.1.3 Simulating impacts on income**

As mentioned, our simulations are, at best, illustrative. Relying on cross-sectional correlations, making important (untestable) assumptions on the law of motion of human capital development, and assuming that the variation induced by experiments would largely be consistent across groups and time are necessary for our exercise. If, as [Cunha and Heckman \(2010\)](#) argue, “skills beget skills and abilities beget abilities” these assumptions are overly restrictive and will bias our estimates downward. If on the other hand, as many economists might find natural, there are diminishing marginal returns to interventions, then the forthcoming estimates are too large.

**Table 3** Variables across life stages

Variable (1)	Dataset (2)	Description (3)
<b>Panel A: circumstances at birth</b>		
Gender	CNLSY, NLSY79	A binary variable indicating if a respondent is male or female
Race	CNLSY, NLSY79	A mutually exclusive and exhaustive set of binary variables indicating whether a respondent is black, Hispanic, white, or other
Maternal educational attainment	CNLSY, NLSY79	A mutually exclusive and exhaustive set of binary variables indicating whether a respondent's mother had not completed high school, graduated from high school, attended some college, or obtained a Bachelor's degree or higher at the time of the respondent's birth
Maternal age (at birth)	CNLSY, NLSY79	The age of a respondent's mother at the time of the respondent's birth
Maternal age (at first birth)	CNLSY, NLSY79	The age of a respondent's mother at the time of the mother's first birth
Marital status of parents	CNLSY	A binary variable indicating if a respondent's mother was married at the time of the respondent's birth
Family income	CNLSY	The income of the respondent's family as a fraction of the federal poverty level for that family at the time of the respondent's birth
Low birth weight	CNLSY	A binary variable indicating if a respondent was 5.5 pounds or less at birth
Mother's AFQT score	CNLSY	The percentile score of a respondent's mother on an unofficial version of the Armed Forces Qualification Test (AFQT). The scores were normalized in 3-month age groups and calculated from the mathematical knowledge, arithmetic reasoning, word knowledge, and paragraph comprehension tests from the Armed Services Vocational Aptitude Battery (ASVAB)
Cognitive stimulation score	CNLSY	Score on the Home Observation Measurement of the Environment (HOME) inventory cognitive stimulation subscale. Scores are taken from the first reported administration of the HOME inventory for each respondent (ages 0–6) and standardized by age at testing
Emotional support score	CNLSY	Score on the home observation measurement of the environment (HOME) inventory emotional support subscale. Scores are taken from the first reported administration of the HOME inventory for each respondent (ages 0–6) and standardized by age at testing

PPVT Score	CNLSY	A respondent's score on the Peabody Picture Vocabulary Test (PPVT). Scores are taken from the first reported administration of the PPVT for each respondent (ages 3–4) and standardized by age at testing
<b>Panel B: early childhood (<math>\approx</math> age 5)</b>		
Math achievement	CNLSY	A respondent's score on the math subtest of the Peabody Individual Achievement Test (PIAT). Scores are taken from tests administered between the ages of 4 and 8 and standardized by age at testing
Reading achievement	CNLSY	A respondent's score on the reading recognition subtest of the PIAT. Scores are taken from tests administered between the ages of 4 and 8 and standardized by age at testing
Antisocial behavior	CNLSY	A respondent's score on the antisocial behavior subscale from the Behavior Problems Index (BPI). Scores are taken from tests administered between the ages of 4 and 8 and standardized by age at testing
Hyperactivity	CNLSY	A respondent's score on the hyperactivity subscale from the BPI. Scores are taken from tests administered between the ages of 4 and 8 and standardized by age at testing
<b>Panel C: middle childhood (<math>\approx</math> age 11)</b>		
Math achievement	CNLSY	A respondent's score on the math subtest of the PIAT. Scores are taken from tests administered between the ages of 9 and 11 and standardized by age at testing
Reading achievement	CNLSY	A respondent's score on the reading recognition subtest of the PIAT. Scores are taken from tests administered between the ages of 9 and 11 and standardized by age at testing
Antisocial behavior	CNLSY	A respondent's score on the antisocial behavior subscale from the BPI. Scores are taken from tests administered between the ages of 9 and 11 and standardized by age at testing
Hyperactivity	CNLSY	A respondent's score on the hyperactivity subscale from the BPI. Scores are taken from tests administered between the ages of 9 and 11 and standardized by age at testing

*Continued*

**Table 3** Variables across life stages—cont'd

Variable (1)	Dataset (2)	Description (3)
<b>Panel D: adolescence (<math>\approx</math> age 13–19)</b>		
High school grad status	CNLSY, NLSY79	A binary variable indicating if a respondent graduated high school by age 19. Note that obtaining a GED does not count as graduating in this analysis.
GPA	CNLSY, NLSY79	A respondent's average GPA in their last year of high school. CNLSY: this variable is reported by the respondent. NLSY79: grades are gathered from official transcripts for all classes a respondent took. We then calculate the average grade for the last year in which more than two graded classes were reported.
Criminal conviction	CNLSY, NLSY79	A binary variable indicating if a respondent was ever convicted or ever on probation before the age of 19.
Teen parent	CNLSY, NLSY79	A binary variable indicating if a respondent reported having a child before the age of 19.
Lives independently from parents	CNLSY, NLSY79	A binary variable indicating if a respondent reported living independently from their parents by the age of 19.
Math achievement	CNLSY, NLSY79	CNLSY: A respondent's score on the math subtest of the PIAT. Scores are taken from tests administered between the ages of 12 and 16 and standardized by age at testing. NLSY79: A respondent's raw score on the arithmetic reasoning test from the ASVAB. Scores are taken from tests administered between the ages of 15 and 23 and standardized by age at testing.
Reading achievement	CNLSY, NLSY79	A respondent's score on the reading recognition subtest of the PIAT. Scores are taken from tests administered between the ages of 12 and 16 and standardized by age at testing. NLSY79: A respondent's raw score on the Word Knowledge test from the ASVAB. Scores are taken from tests administered between the ages of 15 and 23 and standardized by age at testing.
Family income	CNLSY, NLSY79	CNLSY: the income of the respondent's family reported between the ages of 12 and 15. NLSY79: the income of the respondent's family reported between the ages of 13 and 22.
Marijuana use	CNLSY, NLSY79	CNLSY: a binary variable indicating if a respondent reported ever using marijuana by age 19. NLSY79: a binary variable indicating if a respondent reported using marijuana in the past year. Responses are taken from surveys administered between the ages of 16 and 23.

Other drug use	CNLSY, NLSY79	CNLSY: a binary variable indicating if a respondent reported yes to “Have you ever used any drugs other than marijuana or amphetamines, such as cocaine, ‘crack’ (‘rock’) cocaine, hallucinogens, downers, sniffing glue, or something else?” NLSY79: a binary variable indicating if a respondent reported using a drug other than marijuana in the past year. Responses are taken from surveys administered between the ages of 15 and 23
Early sex	CNLSY, NLSY79	A binary variable indicating if a respondent reported having sex before the age of 15.
Suspension	CNLSY, NLSY79	CNLSY: a binary variable indicating if a respondent ever reported being suspended or expelled from school. NLSY79: a binary variable indicating if a respondent ever reported being suspended from school.
Fighting	CNLSY, NLSY79	CNLSY: a binary variable indicating if a respondent reported getting in a fight at school in the past year. Responses are taken from surveys administered between the ages of 18 and 20. NLSY79: a binary variable indicating if a respondent reported getting in a fight at school or work in the past year. Responses are taken from surveys administered between the ages of 16 and 23.
Hitting	CNLSY, NLSY79	A binary variable indicating if a respondent reported hitting or seriously threatening to hit someone in the past year. CNLSY: responses are taken from surveys administered between the ages of 18 and 20. NLSY79: responses are taken from surveys administered between the ages of 16 and 23.
Damaging property	CNLSY, NLSY79	A binary variable indicating if a respondent reported intentionally damaging property that did not belong to them in the past year. CNLSY: responses are taken from surveys administered between the ages of 18 and 20. NLSY79: responses are taken from surveys administered between the ages of 16 and 23.
Self-Esteem Index	CNLSY, NLSY79	A respondent’s score on the Rosenberg Self-Esteem Scale. Raw scores are calculated from responses to the 10 Rosenberg Self-Esteem questions on tests administered between the ages of 16 and 19 and then standardized by age at testing.

*Continued*

**Table 3** Variables across life stages—cont'd

Variable (1)	Dataset (2)	Description (3)
Religious service attendance	CNLSY, NLSY79	A categorical variable indicating the frequency a respondent attends religious services. 0 = Not at all, 1 = several times a year or less, 2 = about once a month, 3 = two or three times a month, 4 = about once a week, 5 = More than once a week. CNLSY: responses are taken from surveys administered between the ages of 19 and 20. NLSY79: responses are taken from surveys administered between the ages of 17 and 22.
Gender role attitudes	CNLSY, NLSY79	The average score across five questions on how respondents view women. CNLSY: responses are taken from surveys administered between the ages of 16 and 20. NLSY79: responses are taken from surveys administered between the ages of 17 and 22.
School clubs	CNLSY, NLSY79	A binary variable indicating if a respondent reported belonging to any clubs, teams, or activities in high school. CNLSY: responses are taken from surveys administered between the ages of 15 and 19. NLSY79: responses are taken from surveys administered between the ages of 19 and 27.

**Panel E: transition to adulthood ( $\approx$  age 29)**

Family income	NLSY79	The income of the respondent's family reported between the ages of 27 and 31
College completion	NLSY79	A binary variable indicating if a respondent received a received a Bachelor's degree or higher by age 29
Lives independently from parents	NLSY79	A binary variable indicating if a respondent reported living independently from their parents. Responses are taken from surveys administered between the ages of 26 and 31.

**Panel F: adulthood ( $\approx$  age 40)**

Family income	NLSY79	The income of the respondent's family reported between the ages of 39 and 44.
---------------	--------	---

Notes: This table presents the variables that are included in the each life-stage of our adaptation of the Social Genome Model. Column (2) reports which datasets a given variable comes from. Variables that are reported as coming from both the CNLSY and the NLSY79 are used as linking variables in the simulation. See [Winship and Owen \(2013\)](#) and Online Appendix B for more information.

If public policy were to implement the most successful math and reading interventions when children are in early childhood, middle childhood, and adolescence, the expected test score increase would be  $1.192\sigma$  in math and  $2.449\sigma$  in reading.<sup>35,36</sup> Using the model, the math impact would translate into a 8.28% increase in income at age 40 and the reading impact would translate into a 25.06% increase.<sup>37</sup> Table 4 presents the average successful impact for each category-life-stage and the estimated impact on income at age 40 if only an intervention with that effect was implemented.

Whether or not the cumulative impact is enough to eliminate racial wage inequality depends on one's ability to "tag" (in the sense of Akerlof, 1978) minorities among other things. We will not hazard a quantitative guess, but qualitatively it seems clear that adhering to the best practices gleaned from the literature on randomized field trials discussed in this chapter would significantly reduce, if not eliminate, much of the gap between racial groups in wages and other important economic and social outcomes.

## 5. CONCLUSION

The review of 196 randomized field experiments designed to increase human capital production unearthed several facts. Early childhood investments, on average, significantly increase achievement. Yet, experiments that attempt to alter the home environment in which children are reared in have shown very little success at increasing student achievement. Among school experiments, high-dosage tutoring and "managed" professional development for teachers have shown to be effective. Ironically, high-dosage tutoring of adolescents seems to be as effective—if not more effective—than early childhood investments. This argues against the growing view that there is a point at which investments in youth are unlikely to yield significant returns (Carniero and Heckman, 2003; Cullen et al., 2013). Lastly, charter schools can be effective avenues of achievement-increasing reform, though the evidence on other market-based approaches such as vouchers or school choice has demonstrated less success.

<sup>35</sup> For the time being, there are no RCTs that estimate effects on children's math and reading abilities during the circumstances at birth life-stage. Potential studies that focused on infants that our search returned were mostly excluded from our meta-analysis for not using standardized math or reading measures.

<sup>36</sup> As to not give too much weight in this exercise to any one study, we approximate the impact of the "most successful" intervention for each life-stage as the average of the three largest statistically significant impacts from each category. If there were no significant studies for a given category-life-stage, we assign an impact of zero. The cumulative impacts stated are the sum of the five averages from the category-life-stages—early childhood, home (middle childhood), school (middle childhood), home (adolescence), and school (adolescence). This simple approximation assumes impacts are linearly additive one-time shocks and experiments are externally valid.

<sup>37</sup> Using the cross-sectional estimates generated by Chetty et al. (2014), the expected income gain at age 28 from a  $1.192\sigma$  increase in standardized math scores is 15.62% and from a  $2.449\sigma$  increase in standardized reading scores is 32.08%.

These facts provide reason for optimism. Through the systematic implementation of randomized field experiments designed to increase human capital of school-aged children, we have substantially increased our knowledge of how to produce human capital and have assembled a canon of best practices. And, in an illustrative simulation exercise, we demonstrate that focusing on what we know has the potential to increase income and reduce racial wage inequality.

The question is: do we have the courage to implement, at scale, human capital policies based on best practices developed from these randomized experiments?

**Table 4** Life-cycle model

		Math		Reading	
		Average impact	Percent change	Average impact	Percent change
		Top three (1)	Income at 40 (2)	Top three (3)	Income at 40 (4)
<b>Panel A: early childhood (<math>\approx</math> age 5)</b>					
Early childhood	$0.413\sigma$	2.27%		$0.973\sigma$	5.58%
<b>Panel B: middle childhood (<math>\approx</math> age 11)</b>					
Home	$0.000\sigma$	0.00%		$0.138\sigma$	1.53%
School	$0.521\sigma$	3.66%		$1.123\sigma$	13.37%
<b>Panel C: adolescence (<math>\approx</math> age 13–19)</b>					
Home	$0.000\sigma$	0.00%		$0.000\sigma$	0.00%
School	$0.258\sigma$	2.10%		$0.215\sigma$	2.84%
<b>Panel D: cumulative</b>					
Cumulative		8.28%			25.06%
Baseline average		\$60,752			\$60,752

Notes: This table reports results from a life-cycle simulation using data from the National Longitudinal Surveys of Youth that follows methods described in [Winship and Owen \(2013\)](#). Panel A reports results for the early childhood life stage. Panel B reports results for the middle childhood life stage. Panel C reports results for the adolescence life stage. Columns (1) and (3) report the average of the three largest statistically significant effect sizes from interventions within the given life stage, category, and subject. If there were less than three significant effect sizes, we either report the average of the one or two significant impacts or report an impact of zero if there were no significant effect sizes. The sample includes all studies found that meet our inclusion restrictions and has impact estimates for the given subject. Note that we use cumulative impacts for each intervention instead of annual impacts. Columns (2) and (4) report the simulated impact that the given increase in test scores (in columns (1) and (3), respectively) at the given life stage would have on an individual's income at age 40. Panel D reports the simulated impact on age-40 income of increasing the scores of an individual in each life stage by all amounts specified in Panels A, B, and C for a given subject. The average age-40 income in the baseline sample is reported at the bottom of the table. See the main text and Online Appendix A for details on our search procedure, inclusion restrictions, and the categories of papers. See the main text and Online Appendix B for details on the life-cycle simulation.

**Table A1** Early childhood  
Study

	Study design	Results
<p>An Evaluation of Curriculum, Setting, and Mentoring on the Performance of Children Enrolled in Pre-Kindergarten (Assel et al., 2007). N schools = 32, N classrooms = 76, N students = 603, ages = 4–5, location = Houston, TX. <b>Treatment groups</b> = two treatment conditions: Condition one administered the <i>Let's Begin with the Letter People</i> curriculum; condition two administered the <i>Doors to Discovery</i> curriculum. The control group continued with the normal curriculum.</p>	<p><b>Treatment defined</b> = Two language and literacy curricula—<i>Let's Begin with the Letter People</i> and <i>Doors to Discovery</i>. Additionally, some of the treatment classroom teachers received monitoring that aided in the implementation of the curricula. <b>Randomization</b> = 26 Title 1 prekindergarten classrooms, 19 universal prekindergarten classrooms, and 31 Head Start center classrooms were identified in Houston, TX across 32 schools sites. School sites were randomly assigned to one of the two curricula or control. The classrooms in sites that were randomly assigned to one of the treatment curricula were further randomized into receiving monitoring or no monitoring.</p>	<p><b>Test score</b> = preschool Language Scale-IV edition: Auditory Comprehension subtest; Expressive Vocabulary test. <b>Regression specification</b> = effect sizes were calculated using the average posttest scores. We report the average impact across site types and outcome measures. <b>Results</b> = the <i>Let's Begin with the Letter People</i> curriculum with mentoring treatment had a <math>-0.055\sigma</math> (<math>0.607</math>) impact on reading test scores. The <i>Let's Begin with the Letter People</i> curriculum with no mentoring treatment had a <math>-0.059\sigma</math> (<math>0.674</math>) impact on reading test scores. The <i>Doors to Discovery</i> curriculum with mentoring treatment had a <math>0.045\sigma</math> (<math>0.605</math>) impact on reading test scores. The <i>Doors to Discovery</i> curriculum with no mentoring treatment had a <math>0.184\sigma</math> (<math>0.597</math>) impact on reading test scores.</p>
<p>Beyond the Pages of a Book: Interactive Reading and Language Development in Preschool Classrooms (Wasik and Bond, 2001). N teachers = 4, N students = 127, age = 4, location = Baltimore, MD. <b>Treatment groups</b> = treatment classrooms incorporated book-reading into their classroom curriculum. Control classrooms continued with their regular curriculum. Ninety-five percent of the sample is eligible for free or reduced lunch and 94% of the sample is African American.</p>	<p><b>Treatment defined</b> = Teachers in the treatment condition were trained in interactive reading techniques designed to teach new vocabulary and prompt a classroom discussion of the material. Books, vocabulary lists, and reading-related activities like arts and crafts were provided. Teachers read approximately two books per week. <b>Randomization</b> = four teachers agreed to the study; half were randomly assigned to treatment.</p>	<p><b>Test score</b> = Peabody Picture Vocabulary Test. <b>Regression specification</b> = Effect sizes were calculated using the average growth between post and pretest scores. <b>Results</b> = Treatment had a <math>0.499\sigma</math> (<math>1.015</math>) impact on reading test scores.</p>

*Continued*

**Table A1** Early childhood—cont'd**Study****Study design****Results**

Children At-Risk for Poor School Readiness: The Effect of an Early Intervention Home Visiting Program on Children and Parents (Necochea, 2007). N families = 52, ages = 3–4. **Treatment groups** = The treatment group participated in the Home Instruction for Parents of Preschool Youngsters (HIPPY) program. The control group received no such intervention. Sample composed entirely of low-income families.

**Treatment defined** = treatment families received a 15-week reading curriculum to implement at home. This curriculum was augmented by approximately seven 30–60 min home visits and eight 2–3 h group meetings, the goal of which was to train treatment mothers ineffective curriculum implementation techniques. **Randomization** = Families were stratified by child's age and preschool enrollment, then randomly assigned to treatment.

**Treatment defined** = the CSRP intervention was a professional development program designed to promote self-regulation skills among low-income preschoolers. Treatment teachers learned how to curb antisocial and dominant behaviors while promoting prosocial behaviors. **Randomization** = Recruitment sites were paired on the basis of similar demographic characteristics and one site from each pairing was randomly assigned to treatment.

**Test score** = Peabody Picture Vocabulary Test; Expressive One-Word Picture Vocabulary Test-Revised. **Regression**

**specification** = Effect sizes were calculated using the means of posttest scores adjusted for pretest scores. We report the average effect size across all outcome measures. **Results** = Treatment had a  $0.159\sigma$  (0.281) impact on reading test scores.

CSRP's Impact on Low-Income Preschoolers' Preacademic Skills: Self-Regulation as a Mediating Mechanism (Raver et al., 2011). N recruitment sites = 18, N classrooms = 35, N students = 543, ages = 3–4. **Treatment groups** = treatment sites implemented the Chicago School Readiness Project (CSRP) for the entire academic year. Control sites received no such intervention. Sample composed entirely of Head Start classrooms.

**Test score** = Peabody Picture Vocabulary Test. **Regression**  
**specification** = Hierarchical linear model (student, classroom, site) controlling for gender, race/ethnicity, primary language, family size, whether the child came from a single-parent household, mother's education, income-to-needs ratio, hours worked by the mother in the previous week, teacher's education, teacher's age, teacher's psychological status, availability of a full-time family worker at the Head Start site, size of the Head Start program, proportion of teachers with a bachelor's degree, proportion of teaching assistants with at least some college education, proportion of families with at least one parent employed, and the proportion of families receiving Temporary Assistance for Needy Families.  
**Results** = treatment had a  $0.34\sigma$  (0.14) impact on reading test scores.

Early Intervention in Low-Birth-Weight Premature Infants: Results Through Age 5 Years From the Infant Health and Development Program ([Brooks-Gunn et al., 1994](#)). N infants = 985, N years = 3. **Treatment groups** = The treatment group received home visits and schooling for three years. The control group received no such intervention. Sample composed entirely of premature infants (born at or before 37 weeks gestational age) weighing under 2500 g at birth.

Educational Effects of the *Tools of the Mind* Curriculum: A Randomized Trial ([Barnett et al., 2008](#)). N teachers = 18, N students = 274, age = 3–4. **Treatment groups** = treatment classroom utilized the *Tools of the Mind* curriculum. The control group continued with their normal curricula.

**Treatment defined** = treatment parents received home visits to provide information on child health and development, as well as social support and management strategies for self-identified problems. Parents received home visits an average of three times per month during the first year, and then an average of 1.5 times per month in the two years to follow. Beginning at age 1, treatment children were expected to attend 4 h of school per day, which reinforced the material introduced in the home visits. **Randomization** = children were stratified by birth weight and randomly assigned to treatment.

**Treatment defined** = the *Tools of the Mind* curriculum focuses on the development of broad foundational skills in reading and mathematics, including self-regulation of social and cognitive behaviors, purposeful recollection, symbolic representation, phonemic awareness, knowledge of letters, familiarity with print, counting, one-to-one correspondence, pattern recognition, and numerical recognition. Teachers in the treatment condition received four days of training prior to the start of the school year. **Randomization** = Teachers were stratified into four groups: teachers with a preschool-grade three license; teachers with a K-8 license; teachers with an N-8 license; and teachers who transferred from another school within the district. Teachers within these groups were then randomly assigned to treatment.

**Test score** = Peabody Picture Vocabulary Test-Revised.

**Regression specification** = Effect sizes were calculated using posttest means. We report the average annual impact. **Results** = Treatment had a  $0.142\sigma$  (0.042) impact on reading test scores.

**Test Score** = Woodcock-Johnson: Applied Math Problems and Letter Word Identification subtests; Peabody Picture Vocabulary Test; Expressive One-Word Picture Vocabulary Test.

**Regression specification** = two-level hierarchical linear model (student, classroom) controlling for pretest scores, primary language, and age. The average (weighted by number of observations) effect across all outcome measures is reported.

**Results** = treatment had a  $0.105\sigma$  (0.125) impact on reading test scores.

**Table A1** Early childhood—cont'd

Study	Study design	Results
<p>Effective Early Literacy Skill Development for Young Spanish-Speaking English Language Learners: An Experimental Study of Two Methods (<a href="#">Farver et al., 2009</a>). N students = 94, ages = 3–5, location = Los Angeles, CA.</p> <p><b>Treatment groups</b> = students in the first treatment group received the literacy express preschool curriculum in English-only (English treatment). Students in the second treatment group received the literacy express curriculum initially in Spanish, but transitioned to English over the course of the intervention (transitional treatment). Control students received the High/Scope curriculum.</p>	<p><b>Treatment defined</b> = both treatment groups utilized the Literacy Express Preschool Curriculum. This curriculum focuses on oral language, emergent literacy, basic math and science, and socioemotional development. It is structured around 10 thematic units that are sequenced in order of complexity and the literacy demands placed upon children. The curriculum lasted for approximately 21 weeks. Students in the English treatment were taught in English for the entire 21 weeks. Students in the transitional treatment were taught in Spanish for the first nine weeks, transitioned to English over the next four weeks, and then taught in English for the remainder of the time.</p> <p><b>Randomization</b> = Balancing for gender, students were randomly assigned to one of the three groups.</p>	<p><b>Test score</b> = Test of Preschool Early Literacy: Definitional vocabulary, Phonological awareness, and Print Knowledge subtests.</p> <p><b>Regression Specification</b> = Effect sizes were calculated using the average growth between pre and posttest scores. We report the average effect size across subtests.</p> <p><b>Results</b> = the English treatment had a <math>0.239\sigma</math> (0.084) impact on reading test scores. The transitional treatment had a <math>0.326\sigma</math> (0.085) impact on reading test scores.</p>
<p>Effects of an Early Literacy Professional Development Intervention on Head Start Teachers and Children (<a href="#">Powell et al., 2010</a>). N teachers = 88, N children = 759, ages = 4–5.</p> <p><b>Treatment groups</b> = the treatment group implemented a professional</p>	<p><b>Treatment defined</b> = Treatment teachers received a 1-semester professional development intervention designed specifically for Head Start teachers. Some teachers received on-site coaching and others received remote coaching. The goal</p>	<p><b>Test score</b> = Peabody Picture Vocabulary Test III: Receptive Language; Woodcock–Johnson III: Letter–Word Identification subtest; concepts about print; Test of Preschool Early Literacy: blending subtest.</p> <p><b>Regression</b></p>

development intervention, and the control group was placed on a wait list to receive the same professional development the following semester.

Effects of Preschool Curriculum Programs on School Readiness ([Preschool Curriculum Evaluation Research Consortium, 2008](#)). N preschools = 208, N classrooms = 315, N children = 2,911, ages = 4–5. **Treatment**  
**groups** = Treatment classrooms implemented one of 14 possible curricula. The control classrooms

of the professional development was to improve teachers' use of evidence-based literacy instruction (data-driven instruction). The intervention comprised of a 2-day workshop followed by expert coaching. The intervention used two different cohorts of teachers and students across two years. **Randomization** = Random assignment occurred at the teacher level and was stratified by whether or not the teacher's classroom was in an urban or not urban area. First, teachers were randomly assigned to an intervention semester (fall or spring) and a participation year (first or second). Next, teachers were randomly assigned to on-site or remote coaching condition.

**Treatment defined** = treatment entailed implementation of one of the following curricula: *Bright Beginnings* (BB), *Creative Curriculum* (CC), *Creative Curriculum with Ladders to Literacy* (CCwL), *Curiosity Corner* (CCorn), *DLM Early Childhood Express supplemented with Open Court Reading Pre-K* (DLM), *Doors to Discovery* (DD), *Early Literacy and*

**specification** = hierarchical linear model analysis (student, classroom, Head Start center) controlling for child race-ethnicity, child gender, and year of participation. **Results** = Treatment had a  $0.16\sigma$  (0.09) impact on reading test scores.

**Test score** = Test of Early Reading Ability; Woodcock–Johnson: Letter-Word Identification, spelling, and applied problems subtests; Peabody Picture Vocabulary Test; Test of Language and Development: Grammatic Understanding subtest; child math assessment: composite score; Preschool Comprehension Test of Phonological and Print Processing;

*Continued*

**Table A1** Early childhood—cont'd**Study****Study design****Results**

<p>continued their usual curricula. Eighty-eight percent of the sample came from low-income families.</p>	<p><i>Learning Model</i> (ELLM), <i>Language-Focused Curriculum</i> (LFC), <i>Let's Begin with the Letter People</i> (LB), <i>Literacy Express</i> (LE), <i>Pre-K Mathematics</i> supplemented with <i>DLM Early Childhood Express Software</i> (Pre-K math), <i>Project Approach</i> (PA), <i>Project Construct</i> (PC), or <i>Ready, Set, Leap!</i> (RSL). <b>Randomization</b> = sample teachers were stratified first by recruitment-site and then either by classroom or preschool (different recruitment sites implemented different stratification criteria) and randomly assigned for treatment. Children were randomly assigned to classes.</p>	<p>Elision subtest. <b>Regression specification</b> = Three-level hierarchical linear model (student, classroom, teacher), controlling for age, gender, race/ethnicity, maternal education, disability status indicator, curriculum, and recruitment-site fixed effects. We reported the average effect across outcome measures. <b>Results</b> = The impacts for each treatment were as follows: BB math = <math>0.182\sigma</math> (0.159) and BB read = <math>0.182\sigma</math> (0.147). CC math = <math>0.102\sigma</math> (0.163) and CC read = <math>0.028\sigma</math> (0.164). CCwL math = <math>0.014\sigma</math> (0.258) and CCwL read = <math>-0.163\sigma</math> (0.267). CCorn math = <math>0.041\sigma</math> (0.183) and CCorn read = <math>0.021\sigma</math> (0.170). DLM math = <math>0.007\sigma</math> (0.138) and DLM read = <math>0.133\sigma</math> (0.158). DD math = <math>0.069\sigma</math> (0.149) and DD read = <math>0.130\sigma</math> (0.208). ELLM math = <math>0.044\sigma</math> (0.181) and ELLM read = <math>0.107\sigma</math> (0.176). LFC math = <math>0.118\sigma</math> (0.141) and LFC read = <math>0.129\sigma</math> (0.159). LB math = <math>0.025\sigma</math> (0.148) and LB read = <math>0.022\sigma</math> (0.207). LE math = <math>0.274\sigma</math> (0.136) and LE</p>
---	---	---

Efficacy of a Direct Instruction Approach to Promote Early Learning ([Salaway, 2008](#)). N students = 61, Age = 3–5.

**Treatment Groups** = Students assigned to treatment classrooms used the *Language for Learning* curriculum. The control group continued with its normal curriculum. The sample was drawn primarily from low-income families.

Evaluation of Child Care Subsidies: Findings from Project Upgrade in Miami ([Layzer et al., 2007](#)). N child Care centers = 180, N students = 1,523, ages = 4,

**Treatment Defined** = The *Language for Learning* curriculum is a form of direct instruction that uses small group and individualized instruction to develop literacy skills. Teachers give numerous, fast-paced presentations with frequent opportunities for child response. Instruction was implemented three days per week. **Randomization** = Children were randomly assigned to treatment or control classrooms.

**Treatment defined** = the *RSL* curriculum uses interactive software to develop oral language development, phonological knowledge, and print knowledge.

$\text{read} = 0.458\sigma$  (0.154). Pre-K math math =  $0.309\sigma$  (0.131) and Pre-K math read =  $0.121\sigma$  (0.166). PA math =  $0.122\sigma$  (0.214) and PA read =  $0.154\sigma$  (0.275). PC math =  $-0.0168\sigma$  (0.133) and PC read =  $-0.050\sigma$  (0.166). RSL math =  $0.087\sigma$  (0.119) and RSL read =  $0.049\sigma$ .

**Test Score** = Kaufman Survey of Early Academic Language Skills: vocabulary, and numbers, letters, and words subtests; Dynamic Indicators of Basic Early Literacy Skills: the initial sounds fluency and letter-naming fluency subtests. **Regression Specification** = Effect sizes for each subtest were calculated using the average growth between post and pretest scores. Effect sizes were averaged by subject to estimate a total math and reading impact. **Results** = Treatment had a  $0.468\sigma$  (0.262) impact on math test scores and a  $0.448\sigma$  (0.262) impact on reading test scores.

**Test score** = Test of Preschool Early Literacy: definitional vocabulary, phonological awareness, and print knowledge subtests. **Regression specification** = Three-level

*Continued*

**Table A1** Early childhood—cont'd**Study****Study design****Results**

<p>location = Miami, FL. <b>Treatment groups</b> = three treatment groups: classrooms in the first group utilized the <i>Ready, Set, Leap! (RSL)</i> curriculum; classrooms in the second group used the <i>Building Early Language and Literacy (BELL)</i> curriculum; classrooms in the third group implemented the <i>Breakthrough to Literacy (BTL)</i> curriculum. The control group continued with their normal curricula.</p> <p>Evaluation of Curricular Approaches to Enhance Preschool Early Literacy Skills (Fischel et al., 2007). N preschools = 6, N classrooms = 35, N students = 507, age = 4, location = Southeastern NY.</p> <p><b>Treatment groups</b> = all classrooms used the <i>High/Scope Education Approach</i> curriculum. Treatment classrooms implemented either the <i>Let's Begin with the Letter People</i> curriculum or the <i>Waterford Early Reading Program</i>.</p>	<p>The <i>BELL</i> curriculum entails two daily 15–20 min lessons designed to promote language proficiency, phonological awareness, shared reading skills, and print awareness. The <i>BTL</i> curriculum builds phonological knowledge through a series of exercises and examinations for one book per week. <b>Randomization</b> = Eligible child care centers were randomly assigned to one of the four groups. To be eligible for the study, a child care center had to have a full class of four-year olds predominantly from families receiving subsidies to pay for child care.</p> <p><b>Treatment defined</b> = the <i>Let's Begin with the Letter People</i> curriculum utilized play-centered instruction to motivate students. It introduced early literacy, math, art, music, science, and social skills via a series of games, songs, and stories. Instruction spanned a three-day period. The <i>Waterford Early Reading Program Level 1</i> used computer software to provide individualized instruction and feedback in letter knowledge, print concepts, vocabulary, and story</p>	<p>hierarchical linear model (student, classroom, randomization block), controlling for age, gender, language spoken at home, classroom-mean pretest score, and dominant language of teacher. Effects are reported for an index of the three subtests. <b>Results</b> = The RSL treatment had a <math>0.507\sigma</math> (<math>0.118</math>) impact on reading test scores. The BELL treatment had a <math>0.061\sigma</math> (<math>0.127</math>) impact on reading test scores. The BTL treatment had a <math>0.544\sigma</math> (<math>0.119</math>) impact on reading test scores.</p> <p><b>Test score</b> = Woodcock–Johnson Revised Tests of Achievement: Letter-Word Identification and Dictation subtests; the Peabody Picture Vocabulary Test III. <b>Regression specification</b> = Effect sizes for each outcome were calculated using the average growth between pre and posttest scores. Effect sizes were averaged to estimate a total impact for each treatment. <b>Results</b> = The <i>Let's Begin with the Letter People</i> curriculum had a <math>0.252\sigma</math> (<math>0.420</math>) impact on</p>
--	--	---

curriculum. Sample drawn entirely from head start programs.

Head Start Impact Study: Final Report (Puma et al., 2010). N children = 4,667, ages = 3–4.

**Treatment groups** = treatment children were offered enrollment for one to two years of the Head Start program. Control children applied to the Head Start program, but were not offered enrollment.

structure. Instruction lasted for 15 min daily. **Randomization** = classrooms randomly assigned to one of three groups.

**Treatment defined** = Head Start provides comprehensive services (preschool education, medical, dental, mental health care, and nutrition services) to low-income children in hopes of boosting their school readiness. Researchers investigate the impacts on two different cohorts, a 3-year-old cohort and a 4-year-old cohort. The 3-year-old cohort was exposed to two years of the Head Start program and the 4-year-old cohort was exposed to just one.

**Randomization** = Researchers first recruited 163 Head Start grantee/ delegate agencies from across the nation. They then stratified eligible Head Start centers in these grantee/ delegate agencies by program and student characteristics and then randomly selected three centers from each grantee/delegate agency (note that small centers were combined

reading test scores. The *Waterford Early Reading Program* curriculum had a  $0.079\sigma$  (0.418) impact on reading test scores.

**Test score** = Peabody Picture Vocabulary Test III; Woodcock –Johnson III: Letter-Word Identification, and applied problems subtests. **Regression**

**specification** = Student outcome regressions control for student pretest scores, gender, age at time of assessment, race/ethnicity, primary language at baseline, number of weeks elapsed between 9/1/2002 and fall testing, primary language spoke at home, primary care giver's age, indicator for if both biological parents live with child, indicator for if biological mother is a recent immigrant, mother's highest level of educational attainment, mother's marital status, and an indicator for if mother gave birth to child as a teenager. **Results** = Winning a lottery to attend Head Start had a  $0.135\sigma$  (0.071) impact on math test

*Continued*

**Table A1** Early childhood—cont'd**Study****Study design****Results**

<p>Longitudinal Results of the Ypsilanti Perry Preschool Project: Final Report (<a href="#">Weikart et al., 1970</a>). N children = 123, location = Ypsilanti, MI</p> <p><b>Treatment groups</b> = treatment students were assigned to an early childhood program lasting from age 3 to age 5 and control students were not.</p>	<p>with nearby centers to create “center groups” that were randomized together as one unit). For the 2002–2003 application process, these centers continued with their typical procedure, reviewing applications and selecting students that they thought would be a good fit. However, the centers selected approximately 40% more students than they had spots for. From the pool of students that each center selected, the researchers then randomly selected students to be offered a spot at that Head Start center.</p> <p><b>Treatment defined</b> = The Perry Preschool Program consisted of children attending 2.5 h of preschool on weekdays during the school year and teachers making weekly home visits. The program practiced an active learning curriculum where students were encouraged to plan, carry out, and reflect on their activities. Participants were drawn from the community served by the Perry Elementary School in Ypsilanti, MI. Families were recruited through surveys, neighborhood referrals, and</p>	<p>scores and a <math>0.188\sigma</math> (0.064) impact on reading test scores.</p> <p><b>Test score</b> = Peabody Picture Vocabulary Test. <b>Regression specification</b> = Effect sizes were calculated from posttest means. We report the average annual impact. <b>Results</b> = Assignment to the Perry Preschool Program had a <math>0.655\sigma</math> (0.162) impact on reading test scores.</p>
--	--	---

door-to-door searches. The study focused on disadvantaged children living in adverse situations. In addition, the study only included students in the IQ range 70–85 and students with mental illness were excluded. The intervention was conducted on five different cohorts in the mid-1960s. **Randomization** = The randomization for this study was as follows: (1) for later cohorts, if a child had an older sibling already participating in the study, they were assigned to the same experimental status as their sibling; (2) the remaining students were ranked by their IQ scores. Odd and even ranked students were then assigned to different groups; (3) some students were manually swapped to balance gender and socioeconomic status between the two groups; (4) a coin was flipped to determine which group would be treated and which group would be control; (5) some children initially assigned to treatment, who had employed mothers, were swapped with control, who had unemployed mothers. This was done because the researchers

---

*Continued*

**Table A1** Early childhood—cont'd**Study****Study design****Results**

National Impact Evaluation of the Comprehensive Child Development Program: Final Report ([St. Pierre et al., 1997](#)). N recruitment Sites = 21, N families = 4410. **Treatment groups** = Treatment group took part in Comprehensive Child Development Program (CCDP) for five years; the control group continued as usual. Sample drawn from families with income below the poverty line.

believed it would be hard for working mothers to participate in weekly home visits with teachers.

**Treatment defined** = the CCDP provides physical, social, emotional, and intellectual support to impoverished families, for the purpose of promoting stable childhood development and economic self-sufficiency among families. Recruited families were either expecting a child or had a child under the age of one. The study analyzed one child per family, termed the “focus child.”

**Randomization** = Rural sites were asked to recruit 180 families; urban sites 360. In order for a family to be eligible, they had to (1) have income below the Federal Poverty guidelines, (2) include a pregnant woman or a child under the age of one, and (3) agree to participate in CCDP activities for five years. Further, each site was instructed to recruit a group of families that were representative (in terms of ethnicity and age of mother) of the low-income population that site served. Recruited families at each site were then randomly assigned to a treatment, a control, or a replacement

**Test score** = Peabody Picture Vocabulary Test. **Regression**

**specification** = Effect sizes were calculated from posttest means. We report the average annual impact.

**Results** = Treatment had a  $0.002\sigma$  ( $0.018$ ) impact on reading test scores.

Parental Incentives and Early Childhood Achievement: A Field Experiment in Chicago Heights ([Fryer et al., 2015a](#)). N families = 257. **Treatment**

**Groups** = Two treatment conditions: Condition one parents received cash rewards for participation in treatment programs; condition two received the same rewards but they were deposited into a trust fund which would be paid upon successful enrollment of their child in college. Control parents did not receive incentives.

Poverty, Early Childhood Education, and Academic Competence: The Abecedarian Experiment ([Ramey and Campbell, 1991](#)). N students = 111,

group. The replacement group was used to replace families that dropped out of the program and were not included in the evaluation.

**Treatment defined** = Treatment parents had the opportunity to attend a “Parent Academy,” in which they learned how to effectively involve themselves with their child’s academic work. Parents also received homework assignments, which asked them to practice the skills learned in the sessions. All treatment programs offered monetary incentive, such that parents could receive up to \$3500 via successful completion of Parent Academy classes and assignments, as well as an additional \$3400 if their child performed well on interim evaluations as well as two large end-of-semester assessments.

**Randomization** = families were randomly assigned to a treatment group or control.

**Treatment defined** = this study investigates the impact of the Abecedarian program on academic test scores. The sample consisted of

**Test score** = Woodcock—Johnson III: Letter-Word Identification, applied problems, spelling, and quantitative concepts subtests; Peabody Picture Vocabulary Test. **Regression specification** = OLS regressions controlling for children’s pretest scores, race, gender, age, and mother’s age. **Results** = The cash treatment had a  $0.150\sigma$  (0.158) impact on math test scores and a  $0.046\sigma$  (0.143) impact on reading test scores. The college treatment had a  $0.224\sigma$  (0.166) impact on math test scores and a  $0.119\sigma$  (0.169) impact on reading test scores.

**Test score** = the math and reading clusters from the Woodcock—Johnson Psycho-Educational Battery, Part 2: tests of academic achievement;

*Continued*

**Table A1** Early childhood—cont'd

Study	Study design	Results
<p>ages = 0–8. <b>Treatment</b>  <b>groups</b> = treatment children received a preschool intervention from infancy until they started school (approximately age 5) and an intervention during the first three years of primary school. Control students continued with the normal curriculum.</p>	<p>healthy infants born to impoverished families in a small, southern town. Treatment children were enrolled in a preschool educational program that operated year round. As an infant, they were exposed to a curriculum that included cognitive and fine motor development, social and self-help skills, language, and gross motor skills. As the children grew older, they moved into a preschool program that placed special emphasis on language development and preliteracy skills. In addition, treatment students went through a 6-week summer transitional classroom experience the summer before kindergarten to best prepare them for the classroom experience. Treatment students continued to receive support for the first three years of school. This support came in the form of a home/school resource teacher (HST). The HST provided parents with activities designed for each child, served as a liaison between the school and parents, and helped families with nonschool related problems that might affect the student's learning.</p>	<p>California Achievement Test: Math and reading subtests. <b>Regression</b>  <b>Specification</b> = Effect sizes were calculated from posttest means. We report the average annual impact across outcome measures.  <b>Results</b> = treatment had an annual impact of <math>0.082\sigma</math> (0.110) impact on math test scores and <math>0.133\sigma</math> (0.115) impact on reading test scores.</p>

Project Breakthrough: A Responsive Environment Field Experiment with Pre-School Children from Public Assistance Families ([Cook County Department of Public Aid, 1969](#)). N students = 184, ages = 3–4.

**Treatment groups** = treatment students received the Edison Responsive Environment (ERE) intervention. Control students continued with the normal curricula.

**Randomization** = eligible infants were matched based on high-risk scores (derived from factors such as maternal and paternal education levels, family income, and parents' marital status) and then one infant from each pair was randomly assigned to treatment and the other to control. Upon entry into kindergarten, children within each group were matched based on their 48-month IQ score and then the pairs were randomly split into the new treatment and control groups.

**Treatment defined** = The ERE intervention gave students talking typewriters, which allowed students to select letters of the alphabet at will, and have the typewriter voice their selection. When the child demonstrated sufficient mastery of this machine, the typewriter would then ask children to select a specific letter. The typewriter was available everyday in class. **Randomization** = students were paired by IQ and assigned randomly to either treatment or control. Note that the researchers had another parallel experiment where all students received intensive

**Test score** = Peabody Picture Vocabulary Test. **Regression specification** = effect sizes were calculated using the mean posttest scores. **Results** = treatment had a  $0.448\sigma$  (0.253) impact on reading test scores.

*Continued*

**Table A1** Early childhood—cont'd**Study****Study design****Results**

<p>Promoting Academic and Social-Emotional School Readiness: The Head Start REDI Program (<a href="#">Bierman et al., 2008</a>). N classrooms = 44, N children = 356, age = 4. <b>Treatment groups</b> = the treatment group administered the Head Start REDI program, and the control group continued with the normal Head Start curriculum.</p>	<p>social work services and were also randomly assigned to the ERE intervention or a control group that only received the intensive social work services. However, these two groups of students were randomized at a different time than the two groups that received normal levels of social work services and therefore we cannot directly compare them. For this reason, we only focus on students that received normal levels of social work services.</p> <p><b>Treatment defined</b> = the intervention involved brief lessons, hands on extension activities and specific teaching strategies linked empirically with the promotion of both social-emotional competencies as well as language development and emergent literacy skills. Take-home materials were provided to parents to enhance skill development at home. The study included two cohorts of 4-year-old children that were recruited across two years. <b>Randomization</b> = Forty-four Head Start classrooms were either randomized into an</p>	<p><b>Test score</b> = Expressive One-Word Picture Vocabulary Test; Test of Language Development: Grammatical Understanding and Sentence Imitation subtests; Test of Preschool Early Literacy: Blending, Elision, and print knowledge subtests. <b>Regression specification</b> = two-level hierarchical linear model (child, classroom) controlling for gender, race, site, and cohort. <b>Results</b> = treatment had a <math>0.16\sigma</math> (<math>0.10</math>) impact on reading test scores.</p>
---	--	--

Randomized Field Trial of an Early Literacy Curriculum and Institutional Support System (Cosgrove et al., 2006). N recruitment sites = 3, N classrooms = 38, N students = 466, age = 4. **Treatment groups** = treatment group implemented the *Early Literacy and Learning Model* curriculum. Control group continued with normal curriculum. Sample drawn from low-performing elementary schools with at least one pre-K program.

The Early Training Project for Disadvantaged Children: A Report After Five Years (Klaus and Gray, 1968). N students = 61, Ages = 4–5. **Treatment Groups** = Two treatment groups: one group of students attended a 10-week summer program plus weekly meetings whenever school was not in session for three years; the second group received the same intervention for two years. Control students received no such

enriched intervention treatment condition or a usual practice control condition.

**Treatment defined** = the *Early Literacy and Learning Model* curriculum emphasized letter and sound recognition, as well as phonological and print awareness. Students engaged in both large- and small-group exercises, in which they typically experienced conversation, repetition, print exposure, and vocabulary exercises. Treatment teachers attended a five-day summer training session and a 1-h weekly coaching seminar. **Randomization** = schools were stratified by recruitment site and randomly assigned to treatment.

**Treatment Defined** = The intervention was designed to develop attitudes conducive to school success, including achievement motivation, persistence, delayed gratification, and interest in school-like activities. Treatment took place during a ten-week summer school program. Further, treatment students received weekly in-house visits to reinforce these lessons whenever school was not in session.

**Test score** = Test of Early Reading Ability. **Regression specification** = Two-level hierarchical linear model (child, classroom) controlling for pretest scores, age, gender, urbanicity, and whether the teacher had a bachelor's degree. **Results** = the *Early Literacy and Learning Model* curriculum had a  $0.253\sigma$  (0.079) impact on reading test scores.

**Test Score** = Peabody Picture Vocabulary Test. **Regression Specification** = Effect sizes were calculated using average growth between pretest and posttest scores. We report the average annual effect across the two treatment groups. **Results** = Treatment had a  $0.250\sigma$  (0.139) impact on reading test scores.

*Continued*

**Table A1** Early childhood—cont'd**Study****Study design****Results**

<p>intervention. Sample drawn from “culturally deprived” African American families in segregated schools.</p> <p>The Effects of a Language and Literacy Intervention on Head Start Children and Teachers (<a href="#">Wasik et al., 2006</a>). N preschools = 2, N teachers = 16, N students = 207. <b>Treatment groups</b> = treatment teachers received training in book reading and oral language strategies. Control teachers received no such training.</p> <p>The Effects of the Home Instruction Program for Preschool Youngsters (HIPPY) on Children’s School Performance at the End of the Program and One Year Later (<a href="#">Baker et al., 1998</a>). N families = 182, Age = 4. <b>Treatment Groups</b> = Treatment families implemented the HIPPY intervention. Control families received no such intervention.</p>	<p><b>Randomization</b> = Students were randomly assigned to one of the three conditions.</p> <p><b>Treatment defined</b> = treatment teachers were trained in the following classroom strategies: Asking questions—designed to promote classroom discussion of the text, building vocabulary, and making connections—designed to introduce further applications of the target vocabulary. Each treatment teacher also received toys/props related to the text to incorporate into the lesson. <b>Randomization</b> = one preschool was randomly assigned to treatment.</p> <p><b>Treatment Defined</b> = Treatment mothers received a series of books to read to their children along with a set of guided workbooks. Books and workbook activities became successively harder as families progressed through the program. Treatment mothers implemented the HIPPY intervention daily.</p> <p><b>Randomization</b> = Families were randomly assigned to treatment.</p>	<p><b>Test score</b> = Peabody Picture Vocabulary Test; Expressive One-Word Picture Vocabulary Test.</p> <p><b>Regression specification</b> = effect sizes were calculated using the average growth between post and pretest scores. <b>Results</b> = treatment had a <math>0.549\sigma</math> (1.442) impact on reading test scores.</p> <p><b>Test Score</b> = Metropolitan Readiness Test: Math and Reading subtests.</p> <p><b>Regression Specification</b> = Effect sizes were calculated using posttest scores adjusted for age, gender, family structure, and pretest scores, as well as the parent’s race/ethnicity, education, and public-assistance status. We report the annual impact of the program.</p> <p><b>Results</b> = Treatment had a <math>0.133\sigma</math> (0.141) impact on math test scores and a <math>0.081\sigma</math> (0.141) impact on reading test scores.</p>
--	---	---

Using Television as a Teaching Tool:  
The Impacts of *Ready to Learn*  
Workshops on Parents, Educators, and  
the Children in their Care (Boller  
et al., 2004). N sites = 20, N parents/  
caretakers = 2319. **Treatment**  
**groups** = treatment parents/  
caretakers attended a *Ready to Learn*  
workshop. Control parents/caretakers  
received no such intervention.

**Treatment defined** = the *Ready to Learn Television Service* supported the development of children's educational programming on the Public Broadcasting Service (PBS) and also provided public workshops to teach parents how to involve themselves with the educational content included in these programs. The goal of these programs was to extend those lessons introduced via the television into normal family life.  
**Randomization** = families were randomly assigned to treatment.

**Test score** = Woodcock and Muñoz  
—Sandoval test: Picture Vocabulary  
and Letter-Word Identification  
subtests. **Regression**  
**specification** = OLS regression  
controlling for parental gender, race,  
education, attitudes toward television,  
English ability, whether the family  
lived in a rural area, and prior exposure  
to a *Ready to Learn* workshop, as well as  
the child's age and gender.  
**Results** = treatment had a  $0.023\sigma$   
(0.062) impact on reading test scores.

**Table A2** Home environment**Study****Study design****Results**

A Comparative Study of the Reading Achievement of Second Grade Pupils in Programs Characterized by a Contrasting Degree of Parent Participation (Ryan, 1964). N teachers = 10, N classrooms = 10, N students = 232, grade = 2. **Treatment Groups** = Treatment classrooms incorporated specific parental involvement into their reading programs. The control group incorporated no such intervention. Sample drawn entirely from schools serving primarily middle-class families, as determined by the superintendent.

**Treatment defined** = treatment parents were asked to frequently read at home with their children. Upon completion of a book, students were asked to share the book with the rest of their class via a brief presentation, in which they named the title and author, and then read their favorite passage. Ten minutes of class time were set aside daily for these presentations.

**Randomization** = Classrooms were randomly assigned to treatment. Eight students were dropped at random to balance the treatment and control groups on the basis of sample size, gender distribution, and pretest scores.

**Treatment defined** = Treatment students attended a book fair in the spring of each school year, where they selected 15 books, 12 of which they would receive to read over the summer. Available books fell into four categories: Pop culture, popular book series, culturally relevant, and curriculum relevant.

**Randomization** = A total of 1082 students were randomly assigned to treatment.

**Test score** = Stanford Achievement Test: Paragraph Meaning and Word Meaning subtests. **Regression specification** = Effect sizes were calculated using average growth between pre and posttest scores. We report the average effect size across outcome measures. **Results** = Treatment had a  $0.212\sigma$  (0.634) impact on reading test scores.

Addressing Summer Reading Setback Among Economically Disadvantaged Elementary Students (Allington et al., 2010). N districts = 2, N schools = 17, N students = 1,713, N years = 3. **Treatment groups** = treatment students received books to read over the summer. Control students received no such intervention. Sample drawn from schools with at least 65% of the student body eligible for free or reduced price lunch.

**Test score** = Florida Comprehensive Achievement Test. **Regression specification** = effect size was calculated using the average posttest scores. We report the average annual impact. **Results** = treatment had a  $0.046\sigma$  (0.033) impact on reading test scores.

An Investigation of the Effects of Daily, Thirty-Minute Home Practice Sessions Upon Reading Achievement With Second Year Elementary Pupils (Hirst, 1972). N schools = 2, N classrooms = 2, N students = 96, grade = 2.

**Treatment groups** = treatment students took part in at-home reading instruction with their parents. Control students received no such intervention.

Assessing the Effectiveness of First Step to Success: Are Short-Term Results the First Step to Long-Term Behavioral Improvements? (Sumi et al., 2012). N recruitment Sites = 5, N schools = 48, N teachers = 288, N students = 287, grades = 1–3. **Treatment groups** = treatment students took part in the *First Step* intervention. Control students received no such intervention.

**Treatment defined** = Parents acted as tutors for 30-min after-school reading sessions at home, five days per week. Texts, lesson plans, and tutoring instructions were provided by the school. The intervention included approximately 80 sessions. **Randomization** = students were stratified by gender and then randomly assigned to treatment.

**Treatment defined** = the *First Step* program has three core components: universal screening, classroom instruction, and parental education. Treatment parents and teachers received training from program coaches to learn how to teach students replacement behaviors and properly reward students when these behaviors were used appropriately. The intervention took place in class and at home for approximately 3 months.

**Randomization** = schools were randomly assigned to treatment or control. Participating teachers in all schools then identified students that demonstrated an elevated risk for externalizing school behavior problems using stages 1 and 2 of the Systematic Screening for Behavior Disorders. The three students from each classroom with the highest average scores across three behavioral indices were invited to participate in the study. Eighty-eight percent of invited students obtained consent from their parents and participated in the study.

**Test score** = Gates-MacGinitie Reading Test: Vocabulary and reading comprehension subtests; Stanford Achievement Test: Word Study Skills subtest. **Regression specification** = OLS regression controlling for a quadratic of pretest scores. We report the average effect size across outcome measures.

**Results** = treatment had a  $0.113\sigma$  ( $0.120$ ) impact on reading test scores.

**Test score** = Woodcock-Johnson: Letter-Word Identification subtest.

**Regression specification** = two-level hierarchical linear model (student, classroom) controlling for pretest scores, age, grade, gender, race/ethnicity, free/reduced-lunch eligibility, special education status, self-reported teacher knowledge and skill, as well as scores on the Maladaptive behavior Index.

**Results** = treatment had a  $-0.104\sigma$  ( $0.092$ ) impact on reading test scores.

**Table A2** Home environment—cont'd**Study****Study design****Results**

Collaboration Between Teachers and Parents in Assisting Children's Reading (Tizard et al., 1982). N schools = 6, N students = 1,867, grades = K-2, location = London, UK, N years = 2. **Treatment groups** = two treatment conditions: students in condition one read aloud to their parents at home; students in condition two read aloud to their teachers in school. The control group did not participate.

**Treatment defined** = Parents in condition one agreed to listen to their child as they read aloud and to complete a report card detailing what their child had read. Books were supplied to the student as needed—most children took home an average of two to four books per week. Condition two mirrored condition one, except teachers listened to children read aloud in small groups. **Randomization** = schools were assigned at random to one of the two treatment conditions. Within each school, one classroom was selected at random to receive treatment, while the remainder served as a within school control.

Does Reading During the Summer Build Reading Skills? Evidence from a Randomized Experiment in 463 Classrooms (Guryan et al., 2014). N districts = 7, N schools = 59, N students = 5,319, grades = 2–3, location = NC. **Treatment groups** = the treatment group was given reading comprehension lessons for the summer. The control group received no such

**Treatment defined** = treatment students were given six reading comprehension lessons in the spring that focused on reading activities that would foster engagement with books during the summer. Parents of treatment students were invited to an after-school family literacy event. Treatment students were mailed 10 books, one per week, during the summer. Students were

**Test score** = National Foundation for Education Research Test A: Reading comprehension subtest. **Regression specification** = effect sizes were calculated using posttest means for each school. We report the average annual effect size across schools. **Results** = the home collaboration treatment had a  $0.445\sigma$  (0.906) impact on reading test scores. The teacher help treatment had a  $-0.012\sigma$  (0.930) impact on reading test scores.

**Test score** = Iowa Test of Basic Skills: Reading Comprehension subtest.

**Regression specification** = OLS regressions controlling for pretest reading comprehension test score and classroom fixed effects.

**Results** = treatment had a  $0.014\sigma$  (0.017) impact on reading test scores.

intervention. Sample drawn entirely from North Carolina.

Effect of Early Literacy Intervention on Kindergarten Achievement ([Phillips et al., 1990](#)). N schools = 12, N classrooms = 18, N students = 325, grade = K, location = Canada.

**Treatment groups** = three treatment conditions: students in condition one participated in the *Little Books* intervention at home; students in condition two participated in it in school only; students in condition three participated in the intervention both at home and in school. Control

asked to mail a trifold that included comprehension questions, after they read each book. Students in the control group received no books and participated in six mathematics lessons during the spring while treatment students participated in reading lessons. **Randomization** = student-level randomization stratified by classroom. The teachers teaching the intervention-related treatment and control lessons were also randomly assigned to new classrooms for the teaching portion of the intervention.

**Treatment defined** = students participating in the *Little Books* intervention at home received a new book each week to read with their parents, who in turn received general guidelines of how to help their children read the book. Students participating in the intervention in school received a lesson plan to accompany each book. Those students participating in the intervention both at home and in school would read the same book in both locations, using the parental intervention to reinforce the lessons

**Test score** = Metropolitan Reading Readiness Test. **Regression specification** = effect sizes were calculated using the means of posttest scores. **Results** = the home treatment had a  $0.000\sigma$  (0.816) impact on reading test scores. The school treatment had a  $-0.025\sigma$  (0.817) impact on reading test scores. The home and school treatment had a  $0.337\sigma$  (0.822) impact on reading test scores.

*Continued*

**Table A2** Home environment—cont'd

Study	Study design	Results
<p>classrooms maintained their normal curricula.</p>	<p>of the classroom intervention. Treatment took place in class for 24 weeks. <b>Randomization</b> = schools were grouped into blocks of four based on location (rural, rural collector, urban) and randomly assigned to treatment or control.</p>	
<p>Effects of a voluntary summer reading intervention on reading achievement: Results from a randomized field trial (Kim, 2006). N schools = 10, N students = 552, grades = 3–5. <b>Treatment groups</b> = treatment students received access to eight free books over the summer. Control students were not granted such access.</p>	<p><b>Treatment defined</b> = Treatment students had access to eight free books over their summer vacation. Skill-appropriate texts were selected based on their semantic and syntactic difficulty. Book selection also took into account student reading preferences, which were assessed via a survey. <b>Randomization</b> = to construct the sample, schools were stratified by Title I eligibility and ranked by their percentage of black and Latino students. Researchers then selected the top four Title I schools and the top six nonTitle I schools with the largest percentage of minority students. Students were then stratified by classroom and randomly assigned to treatment.</p>	<p><b>Test score</b> = Iowa Test of Basic Skills; Dynamic Indicators of Basic Early Literacy Skills: Oral Reading Fluency subtest. <b>Regression specification</b> = OLS regression controlling for pretest scores and randomization block. We report the average effect size across outcome measures. <b>Results</b> = Treatment had a <math>0.012\sigma</math> (0.040) impact on reading test scores.</p>
<p>Effects of parent involvement in isolation or in combination with</p>	<p><b>Treatment defined</b> = students in all three groups participated in two</p>	<p><b>Test score</b> = Stanford Diagnostic Mathematics Test III: Computation</p>

peer tutoring on student self-concept and mathematics achievement (Fantuzzo et al., 1995). N students = 72, grades = 4–5, location = Large urban city in northeastern United States.

**Treatment groups** = This study had two treatment groups. One group received a parental involvement intervention (PI) and the other group received both a parental involvement intervention and a reciprocal peer tutoring intervention in mathematics (RPT + PI). The control group received neither intervention.

Evaluation of the First 3 Years of the Fast Track Prevention Trial with Children at High Risk for Adolescent Conduct Problems (Bierman et al., 2002). N recruitment sites = 4, N schools = 54, N classrooms = 401,

45 min math sessions per week for 10 weeks. The control group worked on assignments individually during these sessions with teaching assistants available if necessary. The PI group also worked individually during sessions, but parents of this group would receive regular updates about the level of students' academic effort in the classroom and parent-initiated celebrations of students' achievement were planned. The RPT + PI group followed the same classroom routine and parent involvement intervention, but also received peer tutoring. Students were randomly paired with a tutor and rewards were given to each pairing when team goals were met. **Randomization** = Participants were stratified by pretest scores and then randomly assigned to one of the three groups.

**Treatment defined** = the *Fast Track* program attempts to address school and family risk factors relating to a child's behavior. The hypothesis guiding the program is that improvements in child competencies, parenting

subtest. **Regression**

**specification** = effect sizes were calculated using means adjusted for pretest scores. **Results** = the PI treatment had a  $0.351\sigma$  (0.395) impact on math test scores. The RPT + PI treatment had a  $0.744\sigma$  (0.406) impact on math test scores.

**Test score** = Spache Diagnostic

**Reading Scale. Regression**  
**specification** = ANCOVA with gender, cohort, site, and baseline child and parent demographics as covariates. We report the average annual impact.

*Continued*

**Table A2** Home environment—cont'd

Study	Study design	Results
<p>N students = 891, grades = 1–3, N years = 3. <b>Treatment</b> <b>groups</b> = treatment group participated in the <i>Fast Track</i> intervention program. Control group did not implement such an intervention. Sample drawn from recruitment sites deemed high-risk due to crime and poverty statistics in the surrounding neighborhoods.</p>	<p>effectiveness, school context, and communications between the home and the school will increase gradually over time and lead to a reduction in antisocial behavior. The program content changes each year to keep pace with developmental needs of children and families. This study focuses on the impact of the first three years of the <i>Fast Track</i> program. <b>Randomization</b> = schools were randomly assigned to treatment.</p>	<p><b>Results</b> = Treatment had a <math>0.02\sigma</math> (<math>0.03</math>) impact on reading scores.</p>
<p>Experimental evidence on the effects of home computers on academic achievement among schoolchildren (Fairlie and Robinson, 2013). N districts = 5, N schools = 15, N students = 1,123, grades = 6–10, N years = 2. <b>Treatment</b> <b>groups</b> = treatment students received computers and control students did not.</p>	<p><b>Treatment defined</b> = home computers are provided to treatment students with no strings attached. <b>Randomization</b> = any student who reported not having a home computer at the beginning of the year was eligible for the study. Students were stratified by school and then randomly assigned to treatment or control.</p>	<p><b>Test score</b> = California Standardized Testing and Reporting program. <b>Regression specification</b> = OLS regression controlling for sampling strata, school year, and first quarter grades. <b>Results</b> = treatment had a <math>-0.06\sigma</math> (<math>0.05</math>) impact on math test scores and a <math>-0.05\sigma</math> (<math>0.05</math>) impact on reading test scores.</p>
<p>Fostering Development of Reading Skills Through Supplemental Instruction: Results for Hispanic and Non-Hispanic Students (Gunn et al., 2005). N schools = 13, N students = 299, grades = K-3,</p>	<p><b>Treatment defined</b> = reading instruction entailed 30 min of small-group or individual tutoring daily for two years in addition to normal class time. The instruction utilized the <i>Reading Mastery</i> and <i>Corrective Reading</i></p>	<p><b>Test score</b> = Woodcock–Johnson Revised: Letter–Word Identification, Word Attack, Passage Comprehension, and Reading Vocabulary subtests. <b>Regression specification</b> = Effect sizes were</p>

location = OR. **Treatment groups** = students in the treatment group received supplemental reading instruction and a social behavior intervention; parents of treatment students received parenting training. The control group did not receive such instruction.

Getting parents involved: A field experiment in deprived schools (Avvisati et al., 2014). N schools = 34, N classrooms = 183, N families = 970, grade = 6, location = Paris, France.

**Treatment groups** = treatment parents attended meetings learning

curricula, both of which focus on developing fluent word-recognition. Social behavior interventions sought to reduce acting-out behaviors by teaching and reinforcing appropriate classroom behaviors. Parent instruction entailed group sessions, in which parents reviewed successful child interaction and communication strategies. **Randomization** = To be eligible for the study, students had to either perform below grade level on literacy assessments or exhibit aggressive social behaviors. Eligible students were grouped by community, grade, and ethnicity and then paired by reading ability as determined by the pretest. One student from each pairing was randomly assigned to treatment.

**Treatment defined** = Treatment parents attended meetings to learn how to involve themselves with their children's education both at home and at school. Parents attended at least three initial meetings, the last of which took place after receipt of the end-of-

calculated using the average posttest scores for each group. We report the annual impact across all subjects.

**Results** = Treatment had a  $0.170\sigma$  ( $0.133$ ) impact on reading test scores.

**Test score** = district standardized tests.  
**Regression specification** = OLS regressions controlling for school fixed effects. **Results** = treatment had a  $0.020\sigma$  ( $0.060$ ) impact on math test scores and a  $-0.035\sigma$  ( $0.064$ ) impact on reading test scores.

*Continued*

**Table A2** Home environment—cont'd

Study	Study design	Results
<p>how to get involved with their child's education. Control parents were not invited to such meetings.</p> <p>Head Start Children's Entry into Public School: A Report on the National Head Start/Public School Early Childhood Transition Demonstration Study (Ramey et al., 2000). N sites = 31, N schools = 413, N families = 7,515, grades = K-3, N years = 6.</p> <p><b>Treatment groups</b> = treatment group received transition demonstration services. Control group received no such intervention. Sample drawn from families previously enrolled in the <i>Head Start</i> program.</p>	<p>term report card, and taught parents how to interpret and respond to their child's academic performance. After the third session, parents could attend additional meetings related to parenting strategies, use of the school-related internet, or sessions designed for non-French speakers.</p> <p><b>Randomization</b> = Classrooms were stratified by school and randomly assigned to treatment.</p> <p><b>Treatment defined</b> = The Transition Demonstration Project was designed as a comprehensive follow-up to the traditional <i>Head Start</i> program that focused on the environment in which children prepare for school. The program had three main goals: first, preparing schools to meet the needs of children at varying levels of development; second, preparing families to support the continued growth and academic development of their children; and third, preparing entire communities to invest in education for families and children. To meet these goals, local coordinators were granted the</p>	<p><b>Test score</b> = Woodcock–Johnson: Letter–Word Identification, Passage Comprehension, mathematics computation, and applied problems subtests. <b>Regression specification</b> = Effect sizes were calculated using the average posttest scores. <b>Results</b> = Treatment had a <math>-0.015\sigma</math> (0.131) impact on math scores and a <math>-0.018\sigma</math> (0.131) impact on reading scores.</p>

Making Work Pay: Final Report on the Self-Sufficiency Project for Long-Term Welfare Recipients (Michalopoulos et al., 2002). N parents = 5729, location = British Columbia and New Brunswick, Canada. **Treatment groups** = families assigned to treatment were given the opportunity to participate in a welfare program that increased their income. Control families were not offered enrollment into this program.

freedom to adopt measures they deemed appropriate within the context of their communities. The intervention was introduced after children had completed the *Head Start* program and were enrolled in kindergarten. **Randomization** = Within each site, schools were placed into one of two blocks based on size and ethnic composition of their student body. One block from each site was randomly assigned to treatment.

**Treatment defined** = this study investigated the impact of the Self-Sufficiency Project (SSP) on child achievement. SSP offered a temporary earnings supplement for up to three years to individuals in British Columbia and New Brunswick, Canada. To participate, individuals had to be single parents who had been on income assistance for at least one year and left income assistance for full-time work. The supplement was given in addition to earnings from work. Participants continued to receive payouts as long as they stayed employed full-time (up to three-years). For full-time

**Test score** = Peabody Picture Vocabulary Test-Revised.

**Regression specification** = Researchers report mean test scores for each experimental group. We report average annual impact.

**Results** = The SSP had a  $0.036\sigma$  (0.058) impact on reading test scores.

*Continued*

**Table A2** Home environment—cont'd

Study	Study design	Results
National Evaluation of Welfare-to-Work Strategies ( <a href="#">Hamilton et al., 2001</a> ). N recruitment Sites = 7, N families = 2,332, ages = 3–5. <b>Treatment groups</b> = two treatment conditions: Condition one implemented a Labor Force Attachment (LFA) intervention; condition two implemented a human capital development intervention (HCD). The control group implemented no such intervention. Sample composed entirely of welfare recipients.	<p>workers making minimum wage, the supplement would approximately double their income.</p> <p><b>Randomization</b> = from the entire pool of eligible single-parents, the researchers randomly selected a sample to contact, interview, and invite to be part of the SSP study. Individuals from this sample that completed a survey and signed a consent form were then randomly assigned to treatment and control groups.</p> <p><b>Treatment defined</b> = The LFA intervention emphasized rapid job placement, so that treatment subjects gained exposure to the job market and developed workplace habits and skills. The HCD intervention focused on developing education and basic skills prior to job placement, so that treatment subjects were more likely to excel at and keep their jobs.</p> <p><b>Randomization</b> = In four recruitment sites, applicants were randomly assigned to one of the three conditions. In three of the recruitment sites, the program</p>	<p><b>Test score</b> = Woodcock–Johnson-Revised: Broad reading and math subtests. <b>Regression specification</b> = effect sizes were calculated using posttest means adjusted for baseline characteristics. We report annual impacts. <b>Results</b> = The LFA treatment had a <math>0.009\sigma</math> (0.041) impact on math test scores and a <math>0.007\sigma</math> (0.042) impact on reading test scores. The HCD treatment had a <math>0.007\sigma</math> (0.042) impact on math test scores and a <math>-0.001\sigma</math> (0.043) impact on reading test scores.</p>

Neighborhoods and Academic Achievement: Results from the Moving to Opportunity Experiment (Sanbonmatsu et al., 2006). N children = 5,074, ages = 6–20. Location = Boston, Baltimore, Chicago, Los Angeles, and New York. **Treatment groups** = treatment families received one of two types of housing vouchers, “experimental” or “section 8”. Control families did not receive a housing voucher.

Parent Tutoring as a Supplement to Compensatory Education for First Grade Children (Mehran and White, 1988). N students = 76, grade = 1. **Treatment groups** = Mothers of students assigned to the treatment group received tutor training. Mothers of students assigned to the control group

administrators picked their treatment of choice, and applicants were randomly assigned to either treatment or control.

**Treatment defined** = Through a lottery for housing vouchers among families initially living in public housing, Moving to opportunity randomly assigned families into three groups. Families in an “experimental” group received housing vouchers eligible for use in low-poverty neighborhoods. Families in a “section 8” group received traditional housing vouchers without neighborhood restrictions. Families in the control group did not receive a voucher, but were still eligible for public housing.  
**Randomization** = Random lottery.

**Treatment defined** = Tutor training consisted of two 4-h sessions in July, follow-up meetings twice a week during the summer, and follow-up meetings once a month during the school year that provided teaching methods for reading. Parents were then advised to tutor children for 30 min twice a week during the

**Test score** = The Woodcock – Johnson Revised battery of tests.

**Regression specification** = OLS regression of test score on a treatment group assignment indicator and baseline covariates (child demographics, child health problems, child education, adult and household characteristics).

**Results** = the effect of the experimental treatment was  $0.006\sigma$  ( $0.014$ ) on reading test scores and  $-0.002\sigma$  ( $0.013$ ) on math test scores. The effect of the section 8 group was  $0.006\sigma$  ( $0.015$ ) on reading test scores and  $-0.007\sigma$  ( $0.014$ ) on math test scores.

**Test score** = comprehensive test of basic skills; Woodcock–Johnson Psycho-Educational Battery.

**Regression specification** = Effect sizes were calculated using the means of posttest scores adjusted for the covariance of pretest scores. We report the average effect size across outcome measures. **Results** = The

*Continued*

**Table A2** Home environment—cont'd

Study	Study design	Results
<p>received no training. Sample composed entirely of at-risk students as determined by their teacher.</p> <p>Parent tutoring in reading using literature and curriculum materials: impact on student reading achievement (Powell-Smith et al., 2000). N students = 36, grade = 2, location = rural/suburban school in the Pacific Northwest. <b>Treatment groups</b> = two treatment groups: the first received literature-based home tutoring from parents (LB) and the second received curriculum-based home tutoring from parents (CB). The control group received normal classroom instruction. Sample composed of low readers, as determined by their teachers.</p>	<p>school year. This study lasted through April of the school year. <b>Randomization</b> = researchers randomly assigned one of the two lowest scoring students to the treatment group and the other to the control group. Students with the third and fourth lowest scores were similarly assigned, and so on.</p> <p><b>Treatment defined</b> = all treatment parents participated in a single 1 –1.5 h training session. For both treatment groups, parents conducted four 20 min tutoring sessions with students every week for 15 weeks. Parents in the LB treatment group received a list of books to read during the tutoring session. Parents in the CB treatment group received tutoring materials based on the reading text that students received instruction on in the classroom. Students in this group could either review the story read in class or select a new story. <b>Randomization</b> = Each parent/student pair was randomly assigned to one of the three groups.</p>	<p>treatment had a <math>0.164\sigma</math> (0.281) impact on reading test scores.</p> <p><b>Test score</b> = Test of Reading Fluency. <b>Regression specification</b> = ANCOVA was used to analyze treatment effects. <b>Results</b> = The LB treatment had a <math>-0.344\sigma</math> (0.411) impact on reading test scores and the CB treatment had a <math>-0.174\sigma</math> (0.409) impact on reading test scores.</p>

Supporting Families in a High-Risk Setting: Proximal Effects of the SAFEChildren Preventive Intervention (Tolan et al., 2004). N schools = 7, N families = 424, grade = 1, location = inner-city Chicago. **Treatment groups** = treatment families took part in the SAFE Children intervention program for 22 weeks. Control group received no such intervention.

The effects of a negative income tax on school performance: results of an experiment (Maynard and Murnane, 1979). N students = 851, grades = 4–10, location = Gary, in.

**Treatment defined** = The SAFEChildren program is designed to develop a broad support network for children deemed at risk of developing antisocial behaviors due to their neighborhood. Treatment parents met in groups weekly to work on parenting skills, family relationships, understanding and managing developmental and situational challenges, increasing support among parents, engaging the school, and managing neighborhood issues like violence. Treatment students underwent 30-min reading tutoring sessions twice weekly. **Randomization** = In the spring, parents of kindergarten students enrolled in the seven participating schools were invited to participate in the study the following school year. Families that agreed to participate were stratified by their child's kindergarten classroom and 55% were randomly assigned to treatment.

**Treatment defined** = The study investigated the impact of a negative income tax program on child achievement. The major features of this program were that participants

**Test score** = Woodcock–Johnson Diagnostic Reading Battery.

**Regression specification** = Two-level hierarchical linear model (year, child) controlling for family income, parental marriage status, gender, ethnicity, and school fixed effects.

**Results** = researchers found a  $0.188\sigma$  (0.068) effect on reading scores.

**Test score** = Iowa Test of Basic Skills: Reading subtest. **Regression specification** = Researchers used a multiple linear regression model controlling for pretest scores, child

*Continued*

**Table A2** Home environment—cont'd

Study	Study design	Results
<p><b>Treatment groups</b> = families assigned to the treatment group received a negative income tax and the control group families continued with the typical income tax system. All participants were impoverished and black.</p> <p>The Effects of a Voluntary Summer Reading Intervention on Reading Activities and Reading Achievement (Kim, 2007). N students = 331, grades = 1–5.</p> <p><b>Treatment groups</b> = treatment students received books to read over the summer. Control students received no books.</p>	<p>were guaranteed a minimum annual income and there was a benefit reduction rate (the amount by which the negative income tax payment was reduced for each dollar of income a family earned). It was therefore expected that a negative income tax for poor families would decrease parents' employment and increase total family income.</p> <p><b>Randomization</b> = Families were randomly assigned to treatment and control.</p> <p><b>Treatment defined</b> = treatment students received books matched to their personal preferences and reading level. Additionally, upon completion of a book, children were instructed to send a postcard to their teachers answering questions about the text. Children were instructed to read 10 books over the summer.</p> <p><b>Randomization</b> = students were stratified by grade and classroom and then randomly assigned to treatment.</p> <p><b>Treatment defined</b> = Parents received training once a week for 30 min over 10 weeks. Parents</p>	<p>baseline characteristics, family baseline characteristics, and school characteristics. <b>Results</b> = Treatment had a <math>0.045\sigma</math> (0.061) impact on reading test scores.</p> <p><b>Test score</b> = Stanford Achievement Test. <b>Regression specification</b> = effect size was calculated using average growth between pre and posttest scores. <b>Results</b> = treatment had a <math>0.037\sigma</math> (0.120) impact on reading test scores.</p> <p><b>Test score</b> = Dynamic Indicators of Basic Early Learning Skills: Phoneme Segmentation Fluency and Nonsense</p>
<p>The Effects of Training Parents in Teaching Phonemic Awareness on the Phonemic Awareness and Early</p>		

Reading of Struggling Readers ([Warren, 2009](#)). N parents = 10, N students = 10, grades = K-1.

**Treatment groups** = treatment parents received instruction on how to educate their child in phonemic awareness. Control parents received instruction in reading aloud to their children.

The Impact of a Literature-Based Program on Literacy Achievement, Use of Literature, and Attitudes of Children from Minority Backgrounds ([Morrow, 1992](#)). N schools = 2, N classrooms = 9, N students = 166, grade = 2.

**Treatment groups** = Two treatment groups: treatment group one received literature-based instruction in school and participated in a reading-at-home program; treatment group two received just the school-based instruction. The control group

taught their children for 30 min daily over 10 weeks. **Randomization** = eligible parents came from federally subsidized housing and had children who scored in the bottom 20% of the Dynamic Indicators of Basic Early Learning Skills (DIBELS) Letter-Naming Fluency subtest and below 10 initial sounds in the DIBELS Initial Sounds Fluency subtest. 30 potentially eligible parents were identified; 10 enrolled. Half of these parents were randomly assigned to the treatment group.

**Treatment defined** = treatment entailed establishment of the following elements: classroom literacy centers—quiet spaces stocked with roughly five to eight books per child; three teacher-guided literature activities per week, including discussion of past texts and composition of original written work; and an independent reading and writing period three to five times weekly. Further, children in the reading-at-home group read at home at least twice weekly with their parents.

Word Fluency subtests. **Regression specification** = Effect sizes were calculated using the average difference in pre and posttest scores. The effect sizes for the two subtests were averaged together.

**Results** = Treatment had a  $0.233\sigma$  ( $0.639$ ) impact on reading test scores.

**Test score** = California Test of Basic Skills: Language and Reading subtests. **Regression**

**specification** = The effect size was calculated for each subtest using the average growth between posttest and pretest scores. The resulting effect sizes were averaged across subtests.

**Results** = Treatment one had a  $0.251\sigma$  ( $0.820$ ) impact on reading test scores and treatment two had a  $0.046\sigma$  ( $0.817$ ) impact on reading test scores.

*Continued*

**Table A2** Home environment—cont'd

Study	Study design	Results
<p>continued with their regular curricula.</p> <p>The Impact of Parental Training in Methods to Aid Beginning Reading on Reading Achievement and Reading Attitudes of First-Grade Students (Peeples, 1996). N students = 50, grade = 1, location = Madison County, MS.</p> <p><b>Treatment groups</b> = treatment group parents received home enrichment learning program (help) or tutor training, and control group parents received no training.</p>	<p><b>Randomization</b> = Eligible classrooms were randomly assigned to one of the three conditions. To be eligible, classrooms had to meet the following criteria: literature was not an integral part of the reading curriculum, teachers had no previous training from the district in literature-based instruction, and none had well-designed literacy centers.</p> <p><b>Treatment defined</b> = HELP, the parent tutor training, consisted of methods to assist first grade students with beginning reading. The training consisted of a 1-h home visit where researchers discussed the importance of reading, factors associated with reading, and methods to aid reading at home.</p> <p><b>Randomization</b> = students in this study were chosen from a population of 800 students entering the first grade for the first time and who were participating in a beginning reading program in Madison County School District. Fifty students were then chosen randomly and independently from a</p>	<p><b>Test score</b> = Gates-MacGinitie Reading Test. <b>Regression specification</b> = the effect size was calculated using average posttest scores. <b>Results</b> = treatment had a <math>0.949\sigma</math> (0.298) impact on reading test scores.</p>

Towards Reduced Poverty Across Generations: Early Findings from New York City's Conditional Cash Transfer Program ([Riccio et al., 2010](#)). N families = 4750, N students = 11,311, grades = 4, 7, and 9, location = New York city. **Treatment groups** = treatment parents received incentives, while control parents did not. Sample drawn from districts in New York city with families at or below 130% of federal poverty level.

population using [Cohen \(1965\)](#) formula. They were assigned to the control or treatment groups using a computer-generated list of numbers.

**Treatment defined** = Parents in the treatment group were offered a set of 22 incentives ranging from 20 to 600 dollars based on education-focused conditions (e.g., children's school attendance, test scores, attendance at parent–teacher conferences), health-focused conditions (e.g., maintaining health insurance, Going to doctor, dentist), and workforce-focused conditions (e.g., working or being in job training). **Randomization** = random lottery.

**Test score** = New York state tests.

**Regression specification** = OLS controlling for characteristics of families. Standard errors adjusted to account for multiple observations per family. We report the average annual impact. **Results** = treatment had a  $-0.005\sigma$  (0.022) impact on math test scores and a  $0.005\sigma$  (0.023) impact on reading test scores.

**Table A3** Schools**Study****Study design****Results**

A multistate district-level cluster randomized trial of the impact of data-driven reform on reading and mathematics achievement (Carlson et al., 2011). N states = 7, N districts = 59, N schools = 538, N students = 31,110, grades = 3–8. **Treatment groups** = the treatment group received data driven instruction lessons and the control group did not receive any such lessons.

**Treatment defined** = the Johns Hopkins Center for Data-Driven Reform in Education (CDDRE) worked with treatment districts to implement quarterly student benchmark assessments and provide district and school leaders with extensive training on interpreting and using the data to guide reform. Control districts did not receive any training or consultants. Each district received one year of treatment, but treatment was implemented in waves. **Randomization** = the CDDRE contacted state departments of education in seven states—AL, AZ, IN, MS, OH, PA, and TN—and asked them to nominate districts with large numbers of low-performing schools. District officials were contacted and those that agreed were included in the randomization procedure. District officials then identified schools within their district that they would want to include in treatment. Generally, low performing schools were chosen. After this recruitment process, the randomization process occurred at the district level. The randomization was stratified by state and recruitment wave.

**Treatment defined** = the Pathway Project teaches teachers how to integrate cognitive strategy instruction and process writing to develop students' text-based analytical writing abilities. Teachers assigned to a Pathway Project classroom

**Test score** = Various state-administered tests standardized at the state level. **Regression specification** = two-level hierarchical linear model (school, district) controlling for pretest score at the school level and school level demographics as well as district level demographics. **Results** = treatment had a  $0.059\sigma$  (0.029) impact on math test scores. Treatment had a  $0.033\sigma$  (0.020) impact on reading test scores.

A Randomized Experiment of a Cognitive Strategies Approach to Text-Based Analytical Writing for Mainstreamed Latino English Language Learners in Grades 6–12 (Kim

**Test score** = California Standards Test. **Regression specification** = three-level hierarchical linear model (student, classroom, school randomization block)

et al., 2011). N schools = 15, N teachers = 103, N students  $\approx$  3,000, grades = 6–11, location = Santa Ana Unified School District.

**Treatment groups** = treatment teachers were selected to participate in Pathway Project professional development. Control teachers were not.

A Study of Cooperative Learning in Mathematics, Writing, and Reading in the Intermediate Grades: A Focus Upon Achievement, Attitudes, and Self-Esteem by Gender, Race, and Ability Group (Glassman, 1989). N schools = 2, N classrooms = 24, N students = 441, grades = 3–5, location = Bay Shore, NY.

**Treatment groups** = Treatment classrooms incorporated cooperative learning strategies into their curricula. Control classrooms continued with their normal curricula.

attended a mix of full-day and after-school sessions for intensive training and support from Pathway Project developers over the course of a school year (46 total hours of training). Each participating teacher was paid a \$1000 stipend to complete all research activities. Teachers in the control group were given classroom resources and received the Pathway professional development in the third year of the study. **Randomization** = classrooms were assigned to grade-school blocks. Within these blocks, classrooms were randomly assigned to either the Pathway intervention or the control group.

**Treatment defined** = cooperative learning is designed to change student attitudes toward academic success by focusing on group achievement. In treatment classrooms, students were organized into teams of similar ability and completed group assignments in reading, writing, and mathematics following an initial presentation by the teacher. These treatment classes supplanted normal reading, writing, and math courses for the school year. Treatment teachers underwent an 11-week training period prior to implementation of the experiment. **Randomization** = Classes were stratified by grade and matched based on pretest performance. One class from each pairing was assigned at random to treatment.

controlling for pretest scores.

**Results** = treatment had a  $0.046\sigma$  (0.035) impact on reading test scores.

**Test score** = Iowa Test of Basic Skills: Mathematics and Reading subtests. **Regression specification** = posttest scores adjusted for pretest scores were used to calculate effect sizes. **Results** = Treatment had a  $0.011\sigma$  (0.408) impact on math test scores and a  $0.040\sigma$  (0.408) impact on reading test scores.

*Continued*

**Table A3** Schools—cont'd**Study****Study design****Results**

Accountability and Flexibility in Public Schools: Evidence from Boston's Charters and Pilots ([Abdulkadiroglu et al., 2011](#)). N students ≈ 11,000, grades = K–12, location = Boston, MA.

**Treatment groups** = students in the charter treatment won admission lotteries to oversubscribed charter schools. Students in the pilot treatment won admission lotteries to oversubscribed pilot schools. Control students were losers of the respective lotteries.

**Treatment defined** = this study utilizes random admission lotteries to investigate the impacts of charter and pilot schools on students' achievement. A charter school is a public school that operates fairly autonomously within guidelines laid down by its state. Charter schools are generally free to manage day-to-day operations, hire teachers and let them go, choose salary schedules, and make curricular decisions. Pilot schools have some of the independence of charter schools—they determine their own budgets, staffing, curricula, and scheduling. However, pilots remain part of the Boston school district and their teachers are Boston Teachers Union members covered by most contract provisions related to pay and seniority. Pilot schools are subject to external review, but the review process to date appears to be less extensive and structured than the external state charter reviews.

**Randomization** = Student admission lotteries.

**Treatment defined** = this study is intended to estimate the impact of having a TFA teacher. The control group consisted of those teachers who were not a member of the TFA corps at the time of the study or at anytime in the past.

**Randomization** = Within each region,

**Test score** = Massachusetts Comprehensive Assessment System: Mathematics and Reading subtests. **Regression specification** = OLS regressions. Charter school regressions include dummies for (combination of schools applied to) \* (year of application). Pilot school regressions include dummies for (first choice) \* (year of application) \* (walk zone status).

**Results** = Winning a lottery to a charter school had a  $0.337\sigma$  (0.071) impact on math test scores and a  $0.201\sigma$  (0.068) impact on reading test scores. Winning a lottery to a pilot school had a  $-0.026\sigma$  (0.069) impact on math test scores and a  $0.052\sigma$  (0.067) impact on reading test scores.

**Test score** = Iowa Test of Basic Skills: Mathematics and Reading subtests. **Regression Specification** = The intent-to-treat (ITT) estimates are estimated using a nested model with the student-level model

Alternative routes to teaching: The impacts of teach for America (TFA) on student achievement and other outcomes ([Glazerman et al., 2006](#)). N schools = 17, N classrooms = 100, N students = 1,800, grades = 1–5,

regions = Baltimore, Chicago, Compton, Houston, New Orleans, and the Mississippi Delta **treatment groups** = treatment students received teaching from a TFA teacher. Control students received teaching from a nonTFA teacher.

An Evaluation of a Pilot Program in Reading for Culturally Disadvantaged First Grade Students (Bowers, 1972). N schools = 4, N students = 200, grade = 1. **Treatment groups** = treatment classrooms implemented the *Distar* reading program. Control classrooms continued with their normal curricula. Sample drawn from students eligible for Title 1 assistance, and who scored below the 25th percentile on the Metropolitan Reading Readiness Test.

An Evaluation of reading recovery (Center et al., 1995). N schools = 10, N students = 70, grade = K-1, location = New South Wales, Australia. **Treatment groups** = Treatment students participated in Reading Recovery (RR).

schools were randomly selected. Students were stratified by grades within school and randomly assigned to classrooms either taught by a TFA or nonTFA teacher.

**Treatment defined** = the *Distar* program is a professional development program designed to change how teachers help struggling, disadvantaged youth. The program emphasizes a systematic approach to decoding instruction, in which children learn the standard rules of reading. Treatment teachers attended a one-week instructional workshop prior to the start of the school year. **Randomization** = Researchers randomly selected a sample of 50 eligible students from each school. Within each of these samples, students were randomly assigned to treatment. Teachers were randomly assigned to instruct treatment classrooms.

**Treatment defined** = RR is an early intervention for low performing students. It consists of extensive professional development for the teachers and one-on-one 30-min daily lessons to accelerate the literacy learning of these children. **Randomization** = Teachers in the 10 participating schools identified 20

nested in the block-level model. The model controls for student-level characteristics and block fixed-effects.

**Results** = treatment had a  $0.15\sigma$  (0.04) impact on math test scores and  $0.03\sigma$  (0.04) impact on reading test scores.

**Test score** = Gates-MacGinitie Reading Test: Vocabulary and Comprehension subtests.

**Regression specification** = for each outcome measure, effect sizes were calculated using the posttest means adjusted for pretest scores. We report the average effect across all outcome measures. **Results** = Treatment had a  $0.257\sigma$  (0.181) impact on reading test scores.

**Test score** = the Burt Word Reading Test and the Neale Analysis of Reading Ability.

**Regression specification** = for each outcome measure, effect sizes were calculated using the growth between pretest and posttest scores. We report the

*Continued*

**Table A3** Schools—cont'd**Study****Study design****Results**

<p>Control students continued with business as usual.</p> <p>An Evaluation of Teachers Trained Through Different Routes to Certification: Final Report (Constantine et al., 2009). N districts = 20, N schools = 63, N teachers = 174, N students = 2,610, grades = K–5, location = CA, IL, WI, IA, GA, NJ, and TX. <b>Treatment groups</b> = treatment students were taught by AC teachers. Control students were taught by traditional certified (TC) teachers.</p>	<p>students whom they considered to be at the greatest risk of failure. These students were tested using the Clay Diagnostic Survey. The 12 lowest scoring students from each school were randomly assigned to three groups: The treatment group, the control group, or a holding group. The holding group was excluded from the analysis and solely existed to delay the entry of control students into the RR program (when treatment students completed/dropped out of RR, they were replaced by a holding student).</p> <p><b>Treatment defined</b> = TC programs place teachers in classrooms only after they have completed teaching certification requirements while alternatively certified (AC) programs place teachers in schools before they have completed their requirements. <b>Randomization</b> = to be eligible, teachers had to (1) be relative novices (three or fewer years of teaching experience prior to 2004–05, five or fewer years prior to 2005–06); (2) teach in regular classrooms (for example, not in special education classrooms); and (3) deliver both reading and math instruction to all their own students. In the study schools, every grade that contained at least one eligible AC teacher and one eligible TC teacher was included. Students in these study grades were randomly assigned to be in the class of an AC or a TC teacher.</p>	<p>average effect across all outcome measures. <b>Results</b> = the impact of receiving RR was <math>1.582\sigma</math> (0.321) on reading test scores.</p> <p><b>Test score</b> = the California Achievement Test, 5th Edition. <b>Regression specification</b> = OLS regression that controls for student characteristics (pretest scores in all subjects, race, gender, and free/reduced price lunch status), years of teaching experience, and school fixed-effects. <b>Results</b> = AC teachers had an impact of <math>-0.05\sigma</math> (0.032) on math test scores and <math>-0.01\sigma</math> (0.050) on reading test scores.</p>
---	---	--

An Evaluation of the Teacher advancement Program (TAP) in Chicago: Year One Impact Report (Glazerman et al., 2009). N schools = 16, N students = 3,501, grades = K-8, location = Chicago, IL.

**Treatment groups** = treatment schools implemented the TAP. Control schools continued with business as usual.

An Investigation of the Effects of a Comprehensive Reading Intervention on the Beginning Reading Skills of First Graders at Risk for Emotional and Behavioral Disorders (Mooney, 2003). N schools = 7, N students = 47, grade = 1.

**Treatment groups** = treatment students received the *Sound Partners* reading intervention program. The control group received no such

**Treatment defined** = TAP attempts to increase school and teacher quality through incentives. Teachers receive performance bonuses based on value added to student achievement and classroom observations. Principals receive bonuses based on school-wide value added and quality of program implementation. Other school staff receive incentives based on school-wide value added. TAP also includes weekly meetings of teachers and a teacher mentor component. **Randomization** = schools were grouped by readiness to participate and then randomly assigned to treatment or control (schools with higher readiness had a higher probability of selection). Analyses are weighted to account for this. Within each group, constrained minimization was utilized to ensure that schools were balanced across school size, predominant race, and geographic location.

**Treatment defined** = all students followed the district's core curriculum. Treatment students received the *Sound Partners* intervention, which entails approximately 30 min of reading tutoring five times weekly throughout the school year in addition to the normal curriculum. The intervention targets phonological awareness, letter-sound relationships, word identification, text reading, and writing. **Randomization** = students were randomly assigned to treatment.

**Test score** = Illinois Standards Achievement Test: Mathematics and Reading subtests. **Regression specification** = effect sizes were calculated using posttest means adjusted for family poverty, special needs, language, race/ethnicity, grade level, and over normal age for a grade. **Results** = treatment had a  $-0.04\sigma$  (0.06) impact on math test scores and  $-0.04\sigma$  (0.05) impact on reading test scores.

**Test score** = Woodcock Reading Mastery Test-Revised: basic reading skills and reading comprehension subtests; Dynamic Indicators of Basic Early Literacy Skills: Phoneme Segmentation Fluency, Nonsense Word Fluency, and Oral Reading Fluency subtests. **Regression specification** = For each outcome measure, effect sizes were calculated using the

**Table A3** Schools—cont'd**Study****Study design****Results**

<p>intervention. Sample drawn from students at risk of developing emotional and behavioral disorders as determined by both their teachers and a psychological screening test.</p> <p>Are high-quality schools enough to increase achievement among the poor? evidence from the Harlem children's zone (<a href="#">Dobbie and Fryer, 2011</a>). N students <math>\approx</math> 850, grades = K-8, location = New York city. <b>Treatment groups</b> = the treatment group consists of students that won the lottery to attend the Promise Academy charter schools in Harlem Children's Zone. The control group consists of students that applied and did not win the lottery.</p> <p>Assessment Data — Informed Guidance to Individualize Kindergarten Reading Instruction: Findings from a Cluster-Randomized Control</p>	<p><b>Treatment defined</b> = the promise Academy charter schools are “No Excuses” charter schools. They have an extended school day and year, additional classes and tutoring for struggling students, high-quality teachers, and provide free medical, dental, and mental-health services. The schools also provide student incentives for achievement, nutritious cafeterias meals, and support for parents in the form of food baskets, meals, and bus fare. <b>Randomization</b> = Student admission lotteries.</p> <p><b>Treatment defined</b> = the baseline of professional development that both groups received included a researcher-delivered summer day-long workshop on response to intervention approaches and</p>	<p>average growth between posttest and pretest scores. We report the average effect across all outcome measures. <b>Results</b> = treatment had a <math>0.278\sigma</math> (0.299) impact on reading test scores.</p> <p><b>Test score</b> = state math and reading tests. <b>Regression specification</b> = OLS regressions that control for gender, race, free lunch status, grade fixed effects, and year fixed effects. The middle school regressions additionally control for previous test scores in the same subject, special education status in previous grades, and whether the student spoke English as a second language in previous grades. <b>Results</b> = winning the lottery had a <math>0.121\sigma</math> (0.049) impact on math test scores and a <math>0.036\sigma</math> (0.042) impact on reading test scores.</p> <p><b>Test score</b> = AIMSWeb Letter Sound Fluency; Woodcock –Johnson III: Picture Vocabulary, Letter-Word Identification, and Word</p>
---	---	--

Field Trial (Al Otaiba et al., 2011). N schools = 14, N teachers = 44, N students = 556, grade = K.

**Treatment groups** = treatment teachers received Individualized Student Instruction for Kindergarten (ISI-K) and Assessment to Instruction (A2i) training. Control received general professional development that the treatment teachers also received.

Can a Mixed-Method Literacy Intervention Improve the Reading Achievement of Low-Performing Elementary School Students in an After-School Program? Results From a Randomized Controlled Trial of READ 180 Enterprise (Kim et al., 2011). N students = 296, grades = 4–6, location = Southeastern MA.

**Treatment groups** = Treatment students received READ 180 instruction during an after-school program. Control students were assigned to a regular district after-school program.

individualized instruction. ISI-K training was meant to help teachers differentiate classroom reading instruction. The ISI-K intervention supports teachers' ability to use assessment data to inform instructional amounts, types, and groupings. Teachers can use A2i software to analyze students' language and reading scores and determine recommended amounts of instruction. **Randomization** = Fourteen schools were matched on several demographic criteria as well as reading test scores. One school from each matched-pair was then randomly assigned to treatment.

**Treatment defined** = READ 180 uses a combination of teacher-directed instruction, computer-based reading lessons, and independent reading. The program allows for differentiated instruction in each of the components of reading. **Randomization** = The sample of eligible students included children who scored below proficiency on the Massachusetts comprehensive assessment system. Students in this eligible sample who returned consent forms were stratified by school and grade and then randomly assigned into either the treatment or the control group.

Attack subtests; Dynamic Indicators of Basic Early Literacy Skills: Nonsense Word Fluency and Phoneme Segmenting Fluency subtests.

### **Regression**

**specification** = A hierarchical multivariate linear model. We report the average impact across all outcomes.

**Results** = Treatment had an impact of  $0.18\sigma$  (0.10) on reading test scores.

**Test score** = Stanford Achievement Test 10: Reading comprehension, vocabulary, and spelling subtests; Dynamic Indicators of Basic Early Literacy Skills: Oral ;Reading Fluency subtest. **Regression specification** = OLS regression controlling for student characteristics including pretest fluency score and school-grade randomization block fixed-effects. **Results** = Treatment had an impact of  $0.20\sigma$  (0.09) on reading test scores.

**Table A3** Schools—cont'd**Study****Study design****Results**

Can interdistrict choice boost student achievement? the case of Connecticut's interdistrict magnet school program ([Bifulco et al., 2009](#)). N students = 494, grades = 6–8 location = near Hartford, CT. **Treatment groups** = the treatment group consisted of students who won an admission lottery and were assigned to magnet schools. The control group consisted of students who lost the same lottery.

Career Academies: Impacts on students' engagement and performance in high school ([Kemple and Snipes, 2000](#)). N schools = 9, N students = 1,764, grade = 8–9, N years = 4. **Treatment groups** = Treatment students received the *Career Academy* intervention. Control students received no such intervention.

Charter Schools in New York City: Who Enrolls and How They Affect Their Students' Achievement ([Hoxby and](#)

**Treatment Defined** = This study investigates the impact of attending two interdistrict magnet schools. The goal was to promote racial and economic integration by allowing students from different school districts to integrate. One school served grades 6–8, and the other served grades 6–12. **Randomization** = lottery-based admissions to charter schools. Admission lotteries held in each of five districts for each of the two schools.

**Treatment defined** = treatment students remain with the same group of teachers throughout high school to develop stronger, supportive educational relationships. Their curriculum includes both academic lessons and vocational material, while the school builds relationships with local employers to provide students with career and work-based learning opportunities. **Randomization** = students were assigned at random to treatment.

**Treatment defined** = A charter school is a public school that operates fairly autonomously within guidelines laid down by its state. Charter schools are

**Test score** = Connecticut mastery test—eighth grade reading test. **Regression specification** = OLS regression controlling for pretest scores (fall of fourth grade, fall of sixth grade) and individual covariates (age, gender, ethnicity, free lunch eligibility, and special education status). **Results** = The impact of winning the lottery is  $0.037\sigma$  (0.046) on math test scores and  $0.081\sigma$  (0.054) on reading test scores.

**Test score** = National Educational Longitudinal Survey of 1988: Math and Reading Comprehension batteries. **Regression Specification** = Effect sizes were calculated using regression-adjusted posttest means, controlling for background characteristics. We report the annual impact of the program. **Results** = Treatment had a  $0.004\sigma$  (0.019) impact on math test scores and a  $-0.012\sigma$  (0.019) impact on reading test scores.

**Test score** = New York state Examinations. **Regression specification** = OLS regression of student

[Murarka, 2009](#)). N schools = 42, N students = 32,551, grade = 3–8, location = New York city, NY. **Treatment**

**groups** = treatment group is comprised of lottery winners to charter schools while the control group is comprised of lottery losers.

Classroom Assessment for Student Learning: The Impact on Elementary School Mathematics in the Central Region ([Randel et al., 2011](#)). N districts = 32, N schools = 67, N teachers = 409, N students ≈ 4,700, grade = 4, location = CO. **Treatment**

**groups** = Treatment teachers participated in the Classroom Assessment for Student Learning (CASL) professional development program. Control teachers did not.

generally free to manage day-to-day operations, hire teachers and let them go, choose salary schedules, and make curricular decisions. This study looked at charter schools throughout New York city. **Randomization** = Student admission lotteries.

**Treatment defined** = CASL is a self-executing professional development program in which teachers learn from a CASL textbook and use CASL assessments to better understand their students' progress. Intervention schools were given all available CASL material as if they had purchased it and researchers were not involved in the implementation of CASL in any way. Intervention schools also created teams of three to six teachers for support and discussion of the CASL material. Control schools were given \$1000 to eliminate the alternative hypothesis that any impact was the result of the schools receiving more resources. The study consisted of a training year and an implementation year. We only report results from the implementation year.

**Randomization** = researchers invited public schools in Colorado that were of sufficient size to have at least one fourth grade and one fifth grade teacher. For their search, researchers focused on the 55 districts that had greater than six total schools and elementary principals that had

achievement on charter dummy, students' pretreatment covariates, lottery, school and grade fixed effects. We report annual impacts.

**Results** = treatment had a  $0.092\sigma$  (0.016) impact on math test scores and a  $0.039\sigma$  (0.016) impact on reading test scores.

**Test score** = Math scores from the Colorado state test.

### Regression

**specification** = Regressions controlling for school-level pretest scores, student-level pretest scores, randomization block fixed effects, and a student's grade. We report the average annual impact.

**Results** = Treatment had a  $0.096\sigma$  (0.073) impact on math test scores.

**Table A3** Schools—cont'd

Study	Study design	Results
<p>Closing the Achievement Gap: A Structured Approach to Group Counseling (Campbell and Brigman, 2005). N schools = 20, N students = 480, grades = 5–6. <b>Treatment groups</b> = the treatment group took part in group counseling utilizing the <i>Student Success Skills</i> model. The control group did not take part in such counseling. Sample drawn entirely from students scoring between the 25th and 60th percentile on the math and reading subtests of the Florida comprehensive achievement test.</p>	<p>signed up to be on the mid-continent research for education and Learning's mailing list. To randomly assign the schools that volunteered to participate, districts were first assigned to nine blocks. The first six blocks were the six districts that had multiple elementary schools that agreed to participate. The final three blocks consisted of districts that only had one elementary school that agreed to participate. These blocks were grouped based on locale, location in Colorado, and date the schools elected to participate. A random number generator was then used to randomly assign half of the schools within each block to treatment and the other half to control. When there was an odd number of schools in a block, the additional school was assigned to control.</p> <p><b>Treatment defined</b> = Treatment entailed group counseling for 45 min, once a week, for eight weeks, followed by four follow-up sessions over the next 4 months. The goal of the <i>Student Success Skills</i> model was to develop academic, social, and self-management skills.</p> <p><b>Randomization</b> = Students were stratified by school and grade and assigned randomly to treatment.</p>	<p><b>Test score</b> = Florida comprehensive achievement test: Mathematics and reading subtests. <b>Regression specification</b> = Average growth in test scores was used to calculate effect sizes.</p> <p><b>Results</b> = Treatment had a <math>0.490\sigma</math> (0.116) impact on math test scores and a <math>0.238\sigma</math> (0.114) impact on reading test scores.</p>

Combining cooperative learning and individualized instruction: effects on student mathematics achievement, attitudes, and behaviors (Slavin et al., 1984). N schools = 6, N classrooms = 18, N students = 504, grades = 3 –5. **Treatment groups** = treatment schools were assigned to one of two treatment conditions: Condition one implemented Team-Assisted Individualization (TAI) strategies into their curriculum; condition two utilized the same curriculum as the TAI group, but did not implement a team environment. The control group maintained their normal curricula.

Comer's School Development Program in Prince George's County, Maryland: A Theory-Based Evaluation (Cook et al., 1999). N schools = 23, N students  $\approx$  12,000, grades = 7–8, location = Prince George's County, MD. **Treatment groups** = treatment schools implemented the school development program and control schools did not.

**Treatment defined** = all treatment students worked on an individualized curriculum via a series of instruction sheets, worksheets, and final assessments. Students in the TAI group were assigned to four or five member teams, each with a mix of high and low mathematics achievers as determined by the pretest. Teams were reassigned after four weeks. Students asked their teammates for help if necessary. At the end of each week, a team score was calculated by summing the final assessment scores from each team member. Students in the individual condition received the same instructions, worksheets, and final assessments as the TAI group, but did not work in a team setting. Treatment replaced the normal mathematics curriculum.

**Randomization** = schools were assigned at random to one of the three conditions.

**Treatment defined** = the School Development Program has three program structures: the School Planning and Management Team, the Social Support Team, and the Parent Team. These three structures are supposed to work together according to three process principles: 1) adults within the school should cooperate with each other, always putting student needs above their own; 2) the school should operate with a problem-solving rather than a fault-finding orientation; and 3) decisions should be reached by consensus rather than vote. Comer believes that if the program structures operate under these processes, then the

**Test score** = comprehensive test of basic skills. **Regression specification** = effect sizes were calculated using average growth between pretest and posttest scores. **Results** = the TAI treatment had a  $0.109\sigma$  ( $1.001$ ) impact on test scores. The TAI without the team environment had a  $0.102\sigma$  ( $1.001$ ) impact on test scores.

**Test score** = Maryland State Readiness Test: Math score. **Regression specification** = school-level outcome means were adjusted using MANOVA, for pretest scores, unreliability in these scores, average school-level socioeconomic status, enrollment size, and elementary school California Achievement Test scores. We report annual impacts. **Results** = treatment had a  $0.008\sigma$  ( $0.033$ ) impact on math test scores.

**Table A3** Schools—cont'd**Study****Study design****Results**

<p>Comparing Instructional Models for the Literacy Education of High-Risk First Graders (Pinnell et al., 1994). N districts = 10, N schools = 40, N students = 403, grade = 1. <b>Treatment groups</b> = four treatment conditions: Condition one received the <i>Reading Recovery</i> (RR) intervention; condition two received the <i>Reading Skills</i> (RS) intervention; condition three received a direct instruction (DI) intervention; and condition four received a reading and writing (R&amp;W) intervention. Control students</p>	<p>processes will spread within the school, staff will focus on attaining widely shared goals, trust will be shared, and staff will understand and meet children's needs.</p> <p><b>Randomization</b> = twenty one schools were paired up with regards to racial composition and previous years' test scores. Schools within each pair were then randomly assigned to treatment or control using a coin toss. One school was not paired and was assigned randomly to treatment or control. Note that two pilot schools were included in the treatment sample because "no obvious differences were found when they were or were not included in the analyses".</p> <p><b>Treatment defined</b> = students in the RR condition received daily 30-min lessons, in which they read familiar texts, independently read instructional-level texts, analyzed and discussed these texts, composed sentences, and reconstructed cut-up sentences. Students in the RS condition took part in daily 30-min exercises designed to develop independent reading strategies. Students in the DI condition received individualized tutoring in fundamental reading skills. Students in the R&amp;W condition received small-group instruction designed to develop a systematic approach to reading.</p> <p><b>Randomization</b> = researchers selected</p>	<p><b>Test score</b> = The Woodcock Reading Mastery Tests and the Gates-MacGintie Reading Test. <b>Regression specification</b> = Hierarchical linear model (student, school) controlling for pretest scores. <b>Results</b> = The RR treatment had a <math>0.484\sigma</math> (0.218) impact on reading test scores. The RS treatment had a <math>0.154\sigma</math> (0.203) impact on reading test scores. The DI treatment had a <math>0.190\sigma</math> (0.221) impact on reading test scores. The R&amp;W treatment had a <math>0.222\sigma</math> (0.250) impact on reading test scores.</p>
---	--	---

continued with the normal curricula. Sample composed of those students with the lowest test scores within each school.

Direct Instruction in Fourth and Fifth Grade Classrooms (Sloan, 1993). N schools = 7, N teachers = 10, N students = 173, grades = 4–5, location = in. **Treatment groups** = treatment teachers received direct instruction training. Control teachers continued business as usual.

Early College, Early Success: Early College High School Initiative (ECHSI) Impact Study (Berger et al., 2013). N students = 2,458, grades = 9–12. **Treatment groups** = the treatment group consists of students who were offered admission to an early

four schools from each district, one of which employed the RR program prior to the study. This school was automatically assigned to the RR condition. The remaining schools from each district were assigned at random to the other conditions. Each school then offered a pool of 10 students with the worst test scores, and four of these were assigned randomly to the treatment specific to their school. The remaining six students in each pool were considered control.

**Treatment defined** = the professional development focused on instructing teachers how to use instructional and questioning strategies associated with direct instruction. **Randomization** = the researcher first contacted 15 elementary schools across a district in Indiana. Fourteen principals agreed to let the researcher reach out to the fourth and 5th grade teachers in their schools. Ten teachers contacted agreed to participate in the study. Five of these teachers were randomly assigned to treatment and the other five to control.

**Treatment defined** = early colleges partner with colleges and universities to offer all students an opportunity to earn an associates degree or up to two years of college credits toward a bachelors degree during high school at no or low cost to the students. The underlying assumption is that engaging underrepresented students in a rigorous high school curriculum tied

**Test score** = comprehensive test of basic skills. **Regression specification** = effect sizes were calculated using average growth between pretest and posttest scores. **Results** = Treatment had a  $0.090\sigma$  ( $0.633$ ) impact on math test scores and a  $0.090\sigma$  ( $0.633$ ) impact on reading test scores.

**Test score** = standardized state assessment scores in reading and mathematics. **Regression specification** = Two-level hierarchical linear model (student, school) controlling for gender, race, low income, and standardized achievement scores in prior reading and

**Table A3** Schools—cont'd

Study	Study design	Results
<p>college from a lottery and a comparison group that included students who participated in the lottery but were not offered admission.</p>	<p>to the incentive of earning college credit will motivate them and increase their access to additional postsecondary education and credentials after high school. <b>Randomization</b> = random admission lotteries.</p>	<p>mathematics. <b>Results</b> = winning the lottery of an early college had a <math>0.14\sigma</math> (0.04) impact on reading scores and a <math>0.05\sigma</math> (0.04) impact on math scores.</p>
<p>Effect of Technology-Enhanced Continuous Progress Monitoring on Math Achievement (Ysseldyke and Bolt, 2007). N districts = 7, N schools = 8, N classrooms = 80, N students = 1880. <b>Treatment groups</b> = treatment classrooms incorporated the <i>Accelerated Math</i> technology-enhanced progress monitoring system (<i>AM</i>) for a full year. Control classrooms maintained normal curricula. Sample drawn from schools who previously expressed interest in the <i>AM</i> system, but had not yet implemented it.</p>	<p><b>Treatment defined</b> = the <i>AM</i> progress monitoring system tracks student performance via regular mathematics exercises and assessments. The software generates practice exercises tailored to the individual and provides both students and teachers with immediate feedback. Teachers are thus able to adjust their classroom instruction based on individual performance. The treatment was in addition to normal class time.</p> <p><b>Randomization</b> = teachers were stratified by school and grade then assigned at random to treatment. In schools where teachers taught multiple classes, classrooms were stratified by school and grade then assigned randomly to treatment.</p>	<p><b>Test score</b> = The STAR math assessment; Terra Nova tests: Math subtests. <b>Regression specification</b> = OLS regression controlling for pretest scores and school fixed effects. The average effect across both subtests is reported. <b>Results</b> = treatment had an impact of <math>0.215\sigma</math> (0.114) on math test scores.</p>
<p>Effectiveness of Paraeducator-Supplemented Individual Instruction: Beyond Basic Decoding Skills (Vadasy et al., 2007). N schools = 9, N teachers = 26, N students = 46, grades = 2–3. <b>Treatment groups</b> = treatment students received extracurricular reading tutoring administered by</p>	<p><b>Treatment defined</b> = treatment students received 30 min of extracurricular tutoring per day, four days per week, for 15 weeks. These tutoring sessions included 15 min of phonics instruction and 15 min of oral passage reading. Target skills included letter-sound correspondences, decoding, sight word reading, spelling, and phonics generalizations. Paraeducators received</p>	<p><b>Test score</b> = dynamic indicators of basic early literacy: Oral reading fluency subtest; Woodcock reading mastery test-Revised: Word Attack and word identification subtests. <b>Regression specification</b> = for each outcome measure, effect sizes were calculated using posttest means adjusted</p>

paraeducators between October and March. The control group received no such tutoring during this time. Sample drawn from students who scored between the 10th and 37th percentile on the word identification subtest from the Woodcock Reading Mastery Test-Revised.

Effects of a Volunteer Tutoring Model on the Early Literacy Development of Struggling First Grade Students (Pullen et al., 2004). N students = 49, grade = 1, location = FL.

**Treatment groups** = treatment group received volunteer tutoring from January to April and the control group received normal classroom instruction.

3 h of training prior to the study, plus an additional 60–90 min of on-site training. **Randomization** = students were stratified by school and assigned randomly to treatment.

**Treatment defined** = Tutors implemented a tutoring model that included repeated reading of familiar texts, explicit coaching in decoding and word strategy, and reading new books for forty 15-min sessions throughout the term. Each tutor was provided with materials such as a guided lesson plan, checklist, and leveled books to use during sessions. Tutors were university students who were recruited and hired for this study. Tutors typically were education majors with limited tutoring experience. Tutors received 4 h of training and demonstrated mastery of the tutoring model prior to interacting with students.

**Randomization** = students were tested using the measure of invented spelling. Those who scored at or below the 30th percentile on the invented spelling assessment were eligible to participate in the study. Eligible students were randomly assigned to a treatment or control group.

for pretest scores. We report the average effect across all outcome measures.

**Results** = treatment had a  $0.502\sigma$  (0.311) impact on reading test scores.

**Test score** = Woodcock Diagnostic Reading Battery: Letter Word Identification and Word Attack subtests.

**Regression specification** = for each outcome measure, effect sizes were calculated using posttest means. We report the average effect across all outcome measures.

**Results** = treatment had a  $0.626\sigma$  (0.300) impact on reading test scores.

**Table A3** Schools—cont'd

Study	Study design	Results
<p>Effects of Academic Tutoring on the Social Status of Low-Achieving, Socially Rejected Children (Coie and Krehbiel, 1984). N schools = 7, N students = 40, grade = 4, location = Durham, NC.</p> <p><b>Treatment groups</b> = 3 treatment groups: treatment group one received academic skill training; treatment group two received social skill training; treatment group three received both. The control group received no such training.</p>	<p><b>Treatment defined</b> = Academic training entailed a meeting with individual tutors for 45 min twice weekly from October to April. The social skill training entailed pairing a child with a more popular, same-sex peer and coaching the child on positive behaviors before and after the interaction; training took place in class once a week for six weeks. All trainers were undergraduates who were coached by the authors prior to treatment.</p> <p><b>Randomization</b> = all students from the sample schools completed both a sociometric evaluation and the California achievement tests. From these examinations, researchers identified 40 students who scored below the 36th percentile on their reading test scores and who were ranked as unpopular by their peers. Researchers assigned these students randomly to one of the four groups.</p>	<p><b>Test score</b> = the math and reading portions of the California Achievement Tests.</p> <p><b>Regression specification</b> = for each outcome measure, effect sizes were calculated using posttest means adjusted for pretest scores. We report the average effect across all outcome measures.</p>
<p>Effects of intensive reading remediation for second and third graders and a 1-year follow-up (Blachman et al., 2004). N districts = 4, N schools = 11, N students = 89, grades = 2–3.</p> <p><b>Treatment groups</b> = treatment group utilized an intensive reading intervention in place of their traditional remedial</p>	<p><b>Treatment defined</b> = treatment children received 50 min of individual reading tutoring, five days per week, between September and June. These sessions replaced remedial reading instruction that would otherwise have been implemented by the school. <b>Randomization</b> = students were stratified by school, grade, and gender and then randomly assigned to treatment.</p>	<p><b>Results</b> = the academic training treatment had a <math>0.773\sigma</math> (<math>0.464</math>) impact on math test scores and a <math>0.472\sigma</math> (<math>0.453</math>) impact on reading test scores. The social skills training treatment had a <math>0.326\sigma</math> (<math>0.452</math>) impact on math test scores and <math>0.397\sigma</math> (<math>0.452</math>) impact on reading test scores. The combined training treatment had a <math>0.505\sigma</math> (<math>0.454</math>) impact on math test scores and a <math>0.616\sigma</math> (<math>0.458</math>) impact on reading test scores.</p> <p><b>Test score</b> = Woodcock–Johnson Mastery Tests-Revised: Word Identification and Word Mastery subtests; Woodcock-Johnson Psycho-Educational Battery-Revised: Calculation and Applied Problems subtests ; the Gray Oral Reading Tests Third Edition. <b>Regression</b></p>

reading instruction. The control group continued with normal remedial reading instruction. Sample drawn from students with demonstrated difficulty reading as determined by the pretest.

Effects of Peer-Assisted Learning Strategies With and Without Training in Elaborated Help Giving (Fuchs et al., 1999). N classrooms = 24, grades = 2–4. **Treatment groups** = two treatment groups: one received training in elaborated help-giving prior to implementing peer-assisted learning strategies (PALS); the other implemented PALS without such training. The control group continued with its regular curriculum. All classrooms included at least some children with chronic reading difficulties and problematic social behaviors.

**Treatment defined** = The treatment groups incorporated three 35-min PALS sessions per week into their normal reading time. During these sessions, a high-ability student was paired with a low-ability student (ability determined by the teacher), and the pair would undertake three activities: Partner reading, paragraph shrinking/summarization, and prediction relay. Students offered each other feedback and were awarded points for completing reading activities and giving feedback. The team with the highest points was recognized by the class at the end of each week. Those students selected for training in help-giving learned strategies on how to identify reading difficulties and potential solutions.

**Randomization** = Classrooms were stratified by grade and half were randomly assigned to implement the PALS program; among those selected for treatment, half were again stratified by grade and randomly assigned for advanced instruction in help-giving strategies.

**Specification** = For each outcome measure, effect sizes were calculated using posttest means adjusted for pretest scores. We report the average effects for math and reading outcomes. **Results** = treatment had a  $-0.275\sigma$  (0.243) impact on math test scores and a  $0.728\sigma$  (0.249) impact on reading test scores.

**Test score** = Stanford Diagnostic Reading Test: Reading Comprehension subtest. **Regression**

**specification** = effect sizes were calculated for each grade strata and each treatment. We report the average effect across grades for each treatment. **Results** = the PALS treatment had a  $0.749\sigma$  (0.517) on reading test scores and the PALS with help-giving training treatment had a  $0.355\sigma$  (0.504) impact on reading test scores.

*Continued*

**Table A3** Schools—cont'd**Study****Study design****Results**

Effects of Reading Decodable Texts in Supplemental First-Grade Tutoring (Jenkins et al., 2004). N schools = 11, N students = 121, grade = 1. **Treatment groups** = two treatment groups: one attended tutoring that included more decodable texts, the other attended tutoring with less decodable texts. The control group maintained their normal curricula. Sample drawn from students who scored at or below the 25th percentile on the wide range achievement test.

**Treatment defined** = all treatment students attended 30-min tutoring sessions four days per week, for 25 weeks. Each session included the following components: Practicing letter-sound relations, reading decodable words, spelling, reading nondecodable words, and text reading. As time went on, text reading progressively occupied more of each session. The sessions for the group with more-decodable texts included tutoring on storybooks with a higher concentration of words that could be deconstructed from previous phonetic instruction. **Randomization** = students were randomly assigned to the three groups.

**Test score** = Woodcock Reading Mastery Tests-Revised: Word Attack, Word Identification and Passage Comprehension subtests; Wide Range Achievement Test-Revised: Reading subtest; Test of Word Reading Efficiency: Sight Word and Phonetic Vocabulary subtests.

**Regression specification** = for each outcome measure, effect sizes were calculated using posttest means. We report the average effect across all outcome measures.

**Results** = the more decodable treatment had a  $0.646\sigma$  (0.282) impact on reading test scores. The less decodable treatment had a  $0.673\sigma$  (0.279) impact on reading test scores.

**Test score** = comprehensive tests of basic skills: Mathematics Computations and Concepts/Applications subtests.

**Regression specification** = effect sizes were calculated using posttest means for each outcome. We report the average effect size across the two subtests. **Results** = the MMP treatment had a  $0.180\sigma$

Effects of Whole Class, Ability Grouped, and Individualized Instruction on Mathematics Achievement (Slavin and Karweit, 1985). N classrooms = 22, N students = 480, grades = 3–5, location = Hagerstown, MD. **Treatment groups** = three treatment groups: The first received the *Missouri Mathematics*

**Treatment defined** = all teachers in the treatment group received 3 h of training and additional implementation assistance for one of the following curricula: The *MMP* was a whole-class, group-paced curriculum focused on active and effective teaching in the form of frequent questions and feedback; the *AGAT* curriculum divided the class into small groups based on skill – teachers differentiated material and pace between the groups, and also

*Program (MMP)*; the second received the *Ability Grouped Active Teaching (AGAT)* curriculum; the third received the *Team Assisted Individualization (TAI)* curriculum. The control group maintained their normal curricula.

Enhancing First-Grade Children's Mathematical Development with Peer-Assisted Learning Strategies (Fuchs et al., 2002). N teachers = 20, N students = 327, grade = 1.

**Treatment groups** = treatment classrooms incorporated peer-assisted learning strategies (PALS) into their curricula. The control group continued with normal curricula.

Enhancing Kindergarteners' Mathematical Development: Effects of Peer-Assisted Learning Strategies (Fuchs et al., 2001). N schools = 5, N teachers = 20, N students = 228, grade = K.

incorporated frequent question and answer; the *TAI* curriculum had children complete independent work while receiving help from peers of similar skill level, while teachers provided guidance and assistance as necessary.

**Randomization** = teachers were assigned randomly to one of the four groups.

**Treatment defined** = all teachers followed the district's core curriculum. Treatment teachers incorporated PALS exercises in class for 30-min, three times weekly, for 16 weeks. During these exercises, students worked cooperatively on math games officiated by the teacher. Students were paired by mathematics ability for three-week cycles. The stronger student acted first as tutor to the lower-performing student, and these roles switched halfway through the cycle. Teachers reassigned the pairings at the end of each cycle.

**Randomization** = teachers were stratified by school and assigned at random to treatment.

**Treatment defined** = all teachers followed the district's core curriculum. Treatment teachers incorporated PALS exercises in class for 20-min, twice weekly, for 15 weeks. During these exercises, students worked cooperatively on math games

(0.587) impact on math test scores. The AGAT treatment had a  $0.751\sigma$  (0.655) impact on math test scores. The TAI treatment had a  $0.361\sigma$  (0.641) impact on math test scores.

**Test score** = the Stanford achievement test. **Regression specification** = average growth in test scores was used to calculate effect sizes.

**Results** = treatment had an impact of  $0.250\sigma$  (0.449) on math test scores.

*Continued*

**Table A3** Schools—cont'd**Study****Study design****Results**

<p><b>Treatment groups</b> = treatment classrooms incorporated peer-assisted learning strategies (PALS) into their curricula. The control group maintained normal curricula.</p> <p>Enhancing the Efficacy of Teacher Incentives Through Loss Aversion (Fryer et al., 2015b). N schools = 9, N students <math>\approx</math> 2,150, grades = K-8, location = Chicago Heights, IL. <b>Treatment groups</b> = teachers were assigned to control or one of four incentivized treatment groups—“Loss,” “Gain,” “team Loss” and “team Gain.”</p>	<p>officiated by the teacher. Students were paired by mathematics ability for two-week cycles. The stronger student acted first as tutor to the lower-performing student. These roles switched halfway through the cycle. Teachers reassigned the pairings at the end of each cycle.</p> <p><b>Randomization</b> = teachers were stratified by school and assigned at random to treatment.</p> <p><b>Treatment defined</b> = “Loss” teachers are paid a lump sum in advance and asked to give their money back if their students do not improve sufficiently; “Gain” teachers receive financial incentives in the form of bonuses at the end of the year linked to student achievement; “team Loss” teachers are equivalent to “Loss” teachers but their payout is also based on the improvement of students taught by a teammate teacher in the same school; and “team Gain” teachers are equivalent to “Gain” teachers but their payout is also based on the improvement of students taught by a teammate teacher in the same school. The experiment was run two separate years, randomizing teachers each year. Note that the “team Gain” treatment arm does not exist in the second year of the experiment. We report results for the pooled “Loss” and pooled “Gain” treatments. Note that although math,</p>	<p>sizes. <b>Results</b> = treatment had a <math>0.161\sigma</math> (0.451) impact on math test scores.</p> <p><b>Test score</b> = Illinois State Achievement Test; Iowa Test of Basic Skills. <b>Regression specification</b> = OLS regressions with individual level controls (gender, race, free lunch eligibility, limited english proficiency status, special education status and baseline ThinkLink test scores), school fixed effects, and grade fixed effects. <b>Results</b> = the “Loss” treatment had a <math>0.197\sigma</math> (0.071) impact on math test scores. The “Gain” treatment had a <math>0.097\sigma</math> (0.076) inmpact on math test scores.</p>
---	--	---

Evaluation of Experience Corps:  
Student Reading Outcomes  
(Morrow-Howell et al., 2009).  
N students = 881, grades =  
1–3. Location = Boston, MA;  
New York city, NY; and Port  
Arthur, TX. **Treatment**  
**groups** = treatment students  
participated in the experienced  
corps (EC) program and control  
students did not.

reading, and science test scores were incentivized (the latter only for fourth and 7th grade science teachers), the main analysis of the paper focuses on math achievement due to most students having multiple reading teachers and the science sample being so small.

**Randomization** = Participating teachers with one homeroom class for the entire day were randomly assigned to one of the five groups; for teachers with classes throughout the day, each class was randomly assigned to one of the five groups. Note that teachers in the “team treatments” were paired with a teacher in the same school and treatment group who taught similar grades, subjects, and students.

**Treatment defined** = The early childhood program recruits volunteers aged 55 + to mentor and tutor children who are at risk of academic failure. Volunteers receive training focused on literacy and relationship building. Volunteers work with students one-on-one for about 15 h per week. **Randomization** = at the beginning of the school year, all students in need of reading assistance were referred to the experience corps program. All referred students were then randomly assigned to the treatment or control group.

**Test score** = Woodcock –Johnson: Word Attack and Passage comprehension subtests; Peabody Picture vocabulary test. **Regression specification** = for each outcome measure, effect sizes were calculated using posttest means adjusted for pretest scores, gender, site, grade, race, classroom behavior, individualized education program, and limited English proficiency. The average effect

*Continued*

**Table A3** Schools—cont'd

Study	Study design	Results
<p>Evaluation of Quality Teaching for English Learners (QTEL) Professional Development: Final Report (<a href="#">Bos et al., 2012</a>). N districts = 8, N schools = 52, N teachers = 303, N students = 8,720, grades = 6–8.</p> <p><b>Treatment groups</b> = teachers in the treatment group had access to QTEL professional development; the control teachers did not have access.</p>	<p><b>Treatment defined</b> = QTEL is a professional development program that prepares teachers to instill comfort and ease with the English language, rather than focusing on isolated and discrete language skills. The program was originally designed for teachers that taught English as a second language, but is available to all teachers. Participating teachers receive the following instruction: Small conferences with their peers over the summer; one-on-one coaching sessions with QTEL staff four to six times throughout the year; and monthly lesson-design meetings with QTEL staff. We only report results for the cohort that was exposed to three years of the experiment.</p> <p><b>Randomization</b> = schools were stratified by district and assigned at random to the treatment group.</p>	<p><b>Results</b> = treatment had a <math>0.075\sigma</math> (0.067) impact on reading test scores.</p> <p><b>Test score</b> = The California Standards Test for English Language Arts.</p> <p><b>Regression specification</b> = effect sizes were calculated using regression-adjusted posttest means.</p> <p><b>Results</b> = treatment had a <math>0.01\sigma</math> (0.03) impact on reading test scores.</p>
<p>Evaluation of the DC Opportunity Scholarship Program: Final Report (<a href="#">Wolf et al., 2010</a>). N students = 2300.</p> <p><b>Treatment groups</b> = treatment students were offered a private school voucher. Control students applied to the same program but were not offered a voucher.</p>	<p><b>Treatment defined</b> = The District of Columbia Opportunity Scholarship Program (OSP) is the first federally funded private school voucher program in the United States. Students who applied and were selected by the program were given the option to move from a public school to a participating private school of their choice.</p> <p><b>Randomization</b> = students that</p>	<p><b>Test Score</b> = Stanford Achievement Test 9: Mathematics and reading subtests.</p> <p><b>Regression specification</b> = OLS regressions controlling for student pretest scores, if a student attended a school labeled as needing</p>

Evaluation of the Early Start to Emancipation Preparation Tutoring Program in Los Angeles County, CA ([Courtney et al., 2008](#)). N students = 402.

**Treatment groups** = treatment group received tutoring services from the ESTEP program. Control group continued with usual instruction. Note the sample consists of students aged 14 and 15 that are one to three years behind grade-level reading or math skill. All students are in foster care.

applied to the program were randomly selected to receive scholarship offers. Note that the program only conducted a random lottery within a grade band (K-5, 6-8, or 9-12) if that grade band was over-subscribed.

**Treatment defined** = treatment students received an average of 18 h of tutoring in reading or math, up to a maximum of 65 total hours over the two years of the evaluation. The program was also designed to inform students about educational resources available to them, as well as create a mentoring relationship between the tutor and student. Tutors were local community-college students. **Randomization** = students were referred to the study by their emancipation preparation advisor. Students who elected to participate were then assigned randomly into the treatment group.

improvement between 2003 and 2005, student's age at time of application, student's entering grade, gender, race, special needs, mother's education, mother's employment, household income, number of children in household, and the number of months the student's family has lived at its current address. We report the average annual impact. **Results** = winning the lottery had a  $0.004\sigma$  (0.026) impact on math test scores and a  $0.026\sigma$  (0.028) impact on reading test scores.

**Test score** = Woodcock

—Johnson III: Letter word identification, calculation, and Passage comprehension subtests. **Regression**

**specification** = OLS regression controlling for student's baseline scores, gender, race, ethnicity, physical health, mental health, substance abuse, level of social support, whether the student was placed in a group home, whether the student previously ran away from foster care, and the type of foster care.

*Continued*

**Table A3** Schools—cont'd  
**Study**

	Study design	Results
<p>Evaluation of the Effectiveness of the Alabama Math, Science, and Technology Initiative (AMSTI) (<a href="#">Newman et al., 2012</a>). N schools = 82, N teachers ≈ 780, N students ≈ 20,000, grades = 4–8, location = al. <b>Treatment groups</b> = treatment schools implemented AMSTI. Control schools continued as usual.</p>	<p><b>Treatment defined</b> = AMSTI involves comprehensive professional development delivered through a 10-day summer institute and follow-up training during the school year; access to program materials, manipulatives, and technology needed to deliver hands on, inquiry-based instruction, and in-school support by AMSTI lead teachers and site specialists who offer mentoring and coaching for instruction. <b>Randomization</b> = from the eligible schools that applied to the program, researchers made an effort to select a sample that was representative of the population of schools in the regions involved. Pairs of similar schools were selected from the pool of applicants based on similarity in mathematics achievement, the percentage of minority students, and the percentage of students from low-income households. Within each pair, schools were randomly assigned either to the AMSTI condition or to the control condition.</p>	<p><b>Results</b> = treatment had a <math>0.048\sigma</math> (0.048) impact on math test scores and a <math>0.016\sigma</math> (0.045) impact on reading test scores.</p> <p><b>Test score</b> = Stanford Achievement Test.</p> <p><b>Regression specification</b> = Two-level hierarchical linear model (student, school) controlling for pretest score, grade level, racial/ethnic minority status, eligibility for free or reduced-price lunch, proficiency in English, gender, and matched pairs fixed effects. Note that we only report the results of the first year of the experiment because the researchers did not report reading impacts in the second year. <b>Results</b> = AMSTI had a <math>0.05\sigma</math> (0.02) impact on math test scores and a <math>0.06\sigma</math> (0.02) impact on reading test scores.</p>
<p>Evaluation of the i3 Scale-Up of Reading Recovery: Year One Report (<a href="#">May et al., 2013</a>). N schools = 147, N students = 866, grade = 1.</p>	<p><b>Treatment defined</b> = reading Recovery is a short-term intervention designed to help low performing first grade students catch up to their peers. Teachers involved in the reading Recovery program are specifically</p>	<p><b>Test score</b> = The Iowa Test of Basic Skills: Composite Reading score. <b>Regression specification</b> = A three-level hierarchical linear model</p>

**Treatment groups** = treatment students participated in the reading Recovery program. Control students did not.

Experimental estimates of education production functions ([Krueger, 1999](#)). N students = 11,600, grade = K-3. **Treatment groups** = Two treatment conditions: Condition one classrooms reduced their class

trained on how to work with struggling students and to implement the program's instructional approach. During a school year, these teachers spend approximately half of their work day working with the same eight low-performing students.

**Randomization** = 628 schools were enrolled in the i3 scale-up of reading Recovery. These schools were randomly assigned to 3 blocks. One of these blocks was randomly chosen to participate in an RCT of reading Recovery during the 2011–2012 school year. Within each of these schools, a subsample of low-performing students was identified using the observation survey of early literacy. In each school, the eight students with the lowest scores were matched according to pretest scores and English Language Learner status. One student in each pair was randomly assigned to treatment and the other to control.

**Treatment defined** = Condition one classrooms reduced their class size to 13 – 17 students from a normal level of 22 – 25 students. Condition two classrooms included a teacher's assistant to help manage the larger class size.  
**Randomization** = Students were stratified by school and assigned at random

(student, matched-pair, school) controlling for pretest scores.  
**Results** = treatment had a  $0.47\sigma$  (0.05) impact on reading scores.

**Test score** = Stanford Achievement Test: Mathematics and Reading subtests. **Regression specification** = OLS regression controlling for race, gender, free lunch status, teacher's race, teacher's

*Continued*

**Table A3** Schools—cont'd**Study****Study design****Results**

<p>size; condition two classrooms continued with normal class size but included a teacher's assistant. Control classrooms continued with normal class size.</p> <p>Explaining Charter School Effectiveness (<a href="#">Angrist et al., 2011</a>). N schools = 22, N students = 9,141, grades = 4–8 and 10, location = MA.</p> <p><b>Treatment groups</b> = Treatment group comprised of lottery winners to charter schools while the control group comprised of lottery losers.</p>	<p>to one of the three conditions. Each school had at least one classroom for each treatment condition.</p> <p><b>Treatment defined</b> = A charter school is a public school that operates fairly autonomously within guidelines laid down by its state. Charter schools are generally free to manage day-to-day operations, hire teachers and let them go, choose salary schedules, and make curricular decisions. This study looks at charter schools throughout the state of Massachusetts. <b>Randomization</b> = Student admission lotteries.</p>	<p>experience, teacher's education, fraction of classmates in class previous year, average fraction of classmates together previous year, fraction of classmates on free lunch, fraction of classmates who attended kindergarten, current grade, first grade in sample, and school fixed-effects. We only report results for the small classroom treatment, because results for the other treatment were not reported by math and reading.</p> <p><b>Results</b> = Initial assignment to small classes had a <math>0.107\sigma</math> (0.033) impact on math test scores and a <math>0.133\sigma</math> (0.033) impact on reading test scores.</p> <p><b>Test score</b> = Massachusetts comprehensive assessment system. <b>Regression specification</b> = ITT regression of student achievement on baseline demographic characteristics and a dummy variable set representing every combination of charter school lotteries, year and grade effects.</p>
--	---	---

Final reading outcomes of the national randomized field trial of success for all (Borman et al., 2007). N schools = 35, N students = 2,108, grades = K-2, location = 11 states (largely concentrated in urban Midwest locations, such as Chicago and Indianapolis, and in the rural small town south), N years = 3.

**Treatment groups** = treatment schools adopted the success for all model. Control schools did not adopt this model. Sample drawn from high poverty schools and is a majority African American.

Financial incentives and student achievement: evidence from randomized trials (Fryer, 2011). N students  $\approx$  27,000,

**Treatment defined** = The intervention is purchased as a comprehensive package, which includes materials, training, ongoing professional development, and a well-specified “blueprint” for delivering and sustaining the model. Schools that elect to adopt success for all implement a program that organizes resources to attempt to ensure that every child will reach the third grade on time with adequate basic skills and will continue to build on those skills throughout the later elementary grades. **Randomization** = cluster randomized trial, with schools randomized into the treatment or control conditions. Note that control schools actually implemented success for all, but only in grades 3–5. Treatment schools implemented success for all in grades K-2. Comparisons were then made between the treated K-2 students and the untreated K-2 students. Researchers claim that observations for treatment fidelity did not reveal any significant contamination due to this research design.

**Treatment defined** = In Dallas, students were paid to read books. In New York, students were rewarded according to interim assessments. In Chicago, students

**Results** = Treatment had a  $0.201\sigma$  (0.042) impact on math test scores and a  $0.075\sigma$  (0.035) impact on reading test scores.

**Test score** = Woodcock

—Johnson: Word Attack, word identification, and Passage comprehension subtests.

**Regression specification** = Hierarchical linear model (student, school) controlling for average school pretest scores. We report the average annual effect across subtests.

**Results** = treatment had a  $0.090\sigma$  (0.060) impact on reading test scores.

*Continued*

**Table A3** Schools—cont'd

Study	Study design	Results
<p>grades = 2 (Dallas), 4 and 7 (NYC), and 9 (Chicago).</p> <p><b>Treatment groups</b> = In each district, students in the treatment group received monetary incentives for performance in school according to a simple incentive scheme. Control students were not incentivized.</p>	<p>were paid for classroom grades.</p> <p><b>Randomization</b> = School-level randomization where the method employed is called re-randomization—minimize the maximum z-score from an equation regressing preassigned treatments on race, previous year test score, free lunch status and English Language Learner eligibility.</p>	<p>reading and math achievement scores from the previous two years, race, gender, free/reduced lunch eligibility, English language learner status, the percent of black students in the school, the percent of Hispanic students in the school, and the percent of free/reduced lunch students in the school. For Dallas, regressions also include a control for whether the student took the English or Spanish version of the ITBS/Logramos test in the previous year. For Dallas and New York city, regressions also include an indicator for being in special education. For New York city, regressions also include controls for the number of recorded behavioral incidents a student had in the previous year, as well as the number of recorded behavioral incidents the school had in the previous year.</p> <p><b>Results</b> = Paying students to read books had a <math>0.079\sigma</math> (<math>0.086</math>) impact on math test scores and <math>0.012\sigma</math> (<math>0.069</math>) impact on reading test scores. Paying students for performance on standardized</p>

Full-day versus half-day kindergarten: an experimental study (Holmes and McConnell, 1990). N schools = 20, N students = 637. **Treatment groups** = Treatment students attended full-day kindergarten and control students attended half-day kindergarten.

Homework in Arithmetic (Koch, 1965). N teachers = 3, N classrooms = 3, N students = 85, grade = 6. **Treatment groups** = two treatment groups: The first received a longer daily arithmetic assignment, while the second received a shorter daily arithmetic assignment. The

**Treatment defined** = Treatment schools had a full-day kindergarten schedule whereas control schools operated on a half-day kindergarten schedule.

**Randomization** = Ten of the elementary schools in the school system were randomly chosen to be in the treatment group. The randomization was stratified by whether or not the school was a Chapter I (low socioeconomic status) school.

**Treatment defined** = The first treatment group received a daily homework assignment in arithmetic that took approximately 30 min to complete. The second treatment group received a daily homework assignment in arithmetic that took approximately 15 min. All classes used the same arithmetic textbook.

**Randomization** = classes were randomly assigned to one of the three conditions.

tests had a  $0.008\sigma$  (0.041) impact on math test scores and a  $-0.008\sigma$  (0.025) impact on reading test scores. Rewarding 9th graders for their grades had a  $-0.010\sigma$  (0.023) impact on math test scores and a  $-0.006\sigma$  (0.028) impact on reading test scores.

**Test score** = California achievement tests. **Regression specification** = Posttest

means for the treatment and control groups were compared using Students *t*-tests.

**Results** = The treatment had an impact of  $-0.290\sigma$  (0.088) on math test scores and an impact of  $0.112\sigma$  (0.083) on reading test scores.

**Test score** = Iowa test of basic skills: Arithmetic concepts and Arithmetic problem Solving subtests. **Regression specification** = For each outcome measure, effect sizes were calculated using the average growth between posttest and pretest scores. We report the average effect across all outcome measures.

*Continued*

**Table A3** Schools—cont'd

Study	Study design	Results
<p>control group received no arithmetic homework.</p> <p>Impact of eMINTS Professional Development on Student Achievement (Brandt et al., 2013). N schools = 60, N teachers = 191, N students = 3610, grades = 7–8, location = MO, N years = 2–3. <b>Treatment groups</b> = the program consisted of two treatment groups: Teachers in the first treatment group received a two-year professional development program, eMINTS comprehensive. Teachers in the second treatment group received the same professional development program plus Intel teach program, which adds a third year to the original program length. Control teachers did not receive eMINTS or the Intel teach program. Sample drawn from high poverty rural schools.</p>	<p><b>Treatment defined</b> = the eMINTS program is based on inquiry based learning, high quality lesson design, establishing a community of learners, and technology integration. It provides teachers with approximately 240 h of professional development spanning two years and support that includes monthly classroom visits. The eMINTS and Intel teach program combines additional professional development and Intels' suite of Web-based teaching tools to build on what teachers learned in the first two years of the program. <b>Randomization</b> = Participating schools were randomly assigned to one of the three groups. Schools had to meet requirements under Title I or Missouri's historical requirements for Title II.D.</p>	<p><b>Results</b> = the shorter assignment intervention had a <math>0.158\sigma</math> (1.417) impact on math test scores. The longer assignment intervention had a <math>0.261\sigma</math> (1.444) impact on math test scores.</p> <p><b>Test score</b> = Missouri assessment program.</p> <p><b>Regression specification</b> = Two-level hierarchical linear model (student, school) controlling for block fixed-effects, pretest scores, student gender, race, free/reduced-price lunch status, limited English Proficient status, Individualized education program status, teacher gender, years of teaching, and if the teacher had a graduate degree. Note that we report the average impact across both treatment groups due to the researchers not reporting impacts separately until the third year. For similar reasons, we only report results from the first year of implementation.</p> <p><b>Results</b> = the eMINTS treatment had a <math>0.067\sigma</math> (0.044) impact on math test scores and a <math>0.007\sigma</math> (0.047) impact on reading test scores.</p>

Impacts of Comprehensive Teacher Induction: Results from the Second Year of a Randomized Controlled Study

(Isenberg et al., 2009). N districts = 17, N students ≈ 3,000, grades = K-6.

**Treatment groups** = treatment schools implemented a comprehensive induction program for one or two years. Control schools continued as usual.

**Treatment defined** = new teachers at treatment schools were provided induction services by either the Educational Testing Service or the New Teacher Center (districts were able to select which service they wanted to receive). Teachers exposed to the intervention were assigned to a full-time mentor for support and training, offered monthly professional development sessions, opportunities to observe veteran teachers, and a colloquium at the end of the school year. **Randomization** = researchers invited districts that met certain criteria (at least 570 teachers in elementary schools, at least 50% of students eligible for free/reduced-price lunch, in the continental U.S., and no prior exposure to comprehensive induction) to participate in the study. In smaller participating districts, all elementary schools with eligible teachers (K-6 teacher, not in departmentalized middle schools, new to the profession, and not already receiving support) were included in the study. Larger districts could elect to only provide the researchers with a subset of elementary schools. Participating schools within each district were randomly assigned to a treatment or control group using constrained minimization. Within participating

**Test score** = The state math and reading assessments that each district administered.

**Regression specification** = researchers used a two-level linear hierarchical model (student, school) controlling for student-level pretest scores, student gender, student race/ethnicity, special education status, English-language learner status, free/reduced-price lunch status, overage for grade, teacher age, teacher age squared, teacher gender, race/ethnicity, indicator showing if a teacher's race/ethnicity matches that of a majority of students, teacher route into teaching, teacher highest degree, teacher holds a degree in an education-related field, first-year teacher, teacher hired after the school year began, teacher attended a competitive college, teacher held a nonteacher job for five or more years, grade fixed-effects, and district fixed-effects. We report average annual impacts.

**Results** = Treatment had a

*Continued*

**Table A3** Schools—cont'd**Study****Study design****Results**

Improving Students' Reading Comprehension Skills: Effects of Comprehension Instruction and Reciprocal Teaching (Spörer et al., 2009). N schools = 2, N students = 210, Grades = 3–6, Location = Germany.

**Treatment Groups** = Three treatment conditions: condition one utilized traditional reciprocal teaching (RT) strategies; condition two utilized instructor guided reading strategies (IG); and condition three utilized the reciprocal teaching in pairs (RTP) strategies. Control students continued with the normal curriculum.

Information and employee evaluation: evidence from a randomized intervention in public schools (Rockoff et al., 2012). N principals = 223, N students = 1,434, location = New York city.

schools, all eligible teachers were included in the study. Based on the school's willingness to participate, treatment schools were then placed into groups that either received one year or two years of intervention.

**Treatment Defined** = Students in condition one focused on developing the following four reading strategies: summarizing, questioning, clarifying, and predicting. Students in condition two focused on the same reading strategies, but instruction was carried out in small groups of 4 – 6 students led by an instructor. Students in condition three were first taught the four reading strategies, and then practiced them in pairs. All conditions received two, 45-minute lessons per week. **Randomization** = First, one school was randomly assigned to the traditional instruction condition as control group, whereas the other school was assigned to the intervention. Second, students from the treatment school were randomly assigned to treatment groups.

**Treatment defined** = the intervention consisted of giving New York city principals reports detailing the value added of their teachers relative to similar teachers in NYC and training principals on how to use this information. The program was offered to principals in NYC schools

$0.000\sigma$  (0.044) impact on reading test scores and a  $-0.010\sigma$  (0.044) impact on math test scores.

**Test Score** = Diagnostischer Test Deutsch, assessed 12 weeks after the completion of treatment. **Regression Specification** = ANCOVA analysis controlling for pretest scores. **Results** = The RT treatment had a  $0.681\sigma$  (0.203) impact on reading test scores. The RTP treatment had a  $0.282\sigma$  (0.179) impact on reading test scores. The IG treatment had a  $0.159\sigma$  (0.198) impact on reading test scores.

**Test score** = state test in mathematics and reading. **Regression specification** = Researchers estimated the impact of the intervention by regressing student-level achievement gains on an

**Treatment groups** = treatment principals received reports detailing the teacher's value added of all teachers in their school and training on how to interpret this data. Control principals did not have access to these reports and went about business as usual.

Information and Student Achievement: Evidence from a Cellular Phone Experiment (Fryer, 2013a). N students = 1,907, grades = 6–7, location = Oklahoma city public schools. **Treatment groups** = three groups of treatment students were provided with free cellular phones and daily information about the link between human capital and future outcomes via text messages. Treatment one students received a monthly allocation of credits on their

containing any grade in 4–8. Principals had to sign-up and complete a survey to participate. Over 1000 principals were eligible to participate, but only 305 signed up. Out of the 305 that signed up, only 223 completed the necessary survey.

**Randomization** = principals in the study were assigned to blocks by grade configuration of their schools (elementary, middle, and K–8 schools). Principals within each block were then randomly assigned to treatment via a random number.

**Treatment defined** = treatment one—students received a cell phone (preloaded with 300 min) with Daily informational text messages and a fixed allocation of 200 credits on a monthly schedule. Treatment two—students received a cell phone (preloaded with 300 min), received daily informational text messages, and were required to read books and complete quizzes to confirm their additional understanding of those books to receive additional credits on a bi-weekly basis. Treatment three—students received a cell phone (preloaded with 300 min) and Were required to read books and complete quizzes about those

indicator for if the student was in a treatment school, allowing for random effects at the teacher and school level. The main specification reported does not include any covariates, but the researchers showed that adding in student-level or teacher-level covariates does not significantly alter the results. **Results** = treatment had a  $0.028\sigma$  (0.017) impact on math test scores and a  $0.008\sigma$  (0.013) impact on reading test scores.

**Test score** = Oklahoma core curriculum criterion referenced tests. **Regression specification** = An ITT regression controlling for student-level demographics and school fixed effects.

**Results** = Information had a  $-0.027\sigma$  (0.039) impact on math test scores and a  $0.040\sigma$  (0.041) impact on reading test scores. Nonfinancial incentives had a  $-0.023\sigma$  (0.047) impact on math test scores and  $0.023\sigma$  (0.050) impact on reading test scores.

*Continued*

**Table A3** Schools—cont'd**Study****Study design****Results**

<p>phone and received daily informational messages. Treatment two students received daily informational messages and were required to read books and take quizzes to receive additional credits on their phone. Treatment three students were required to read books and take quizzes to receive additional credits on their phone. Control students did not receive a phone, informational messages, or nonfinancial incentives.</p> <p>Injecting charter school best practices into traditional public schools: evidence from field experiments (Fryer, 2014a). N schools = 20, N students = 39,464, grades = K-5, location = Houston, TX.</p> <p><b>Treatment groups</b> = treatment schools implemented a five-pronged intervention; control schools received no such intervention. Sample composed entirely of low-performing schools.</p>	<p>books to receive additional credits on a biweekly schedule. <b>Randomization</b> = 6th and 7th grade students from the 22 eligible schools in Oklahoma city public schools (all schools with 6th and 7th grade students that were not designated alternative education academies) were eligible to participate in the program. Of those 4810 students, 1907 returned consent forms and were randomized into one of the four groups.</p> <p><b>Treatment defined</b> = Treatment schools implemented the following five practices: Increased instructional time; replacing principals and teachers who failed to adequately increase student achievement; implementing daily high-dosage mathematics tutoring for 4th graders; use of data-driven curricula; and fostering a culture of high expectations.</p> <p><b>Randomization</b> = schools were ranked by aggregate reading and math scores on state achievement tests for grades three through five, as well as by Stanford 10 scores for kindergarten through 2nd grade. The bottom two schools were automatically assigned to treatment. The</p>	<p><b>Test score</b> = statewide math and reading assessments developed by the Texas education Agency. <b>Regression specification</b> = OLS regression controlling for student gender, race, free/reduced price lunch status, English language proficiency, special education accommodations, and enrollment in a gifted or talented program, as well as the school-wide composition of student gender, race/ethnicity, free/reduced price lunch,</p>
---	---	--

KIPP Middle Schools: Impacts on Achievement and Other Outcomes (Tuttle et al., 2013). N schools = 10, N students  $\approx$  1,000, grades = 5–8.

**Treatment groups** = the treatment group consists of students that won a lottery to attend one of the knowledge is power program (KIPP) charter schools. The control group consists of students that applied and did not win the lotteries.

remaining 18 schools were placed into pairs based on aggregate scores and one school from each pairing was assigned randomly to treatment.

**Treatment defined** = KIPP is a national network of public charter schools targeting low-income families. The goal of KIPP is to prepare students for college and set them up to succeed in life. Note only 10 of 53 KIPP middle schools could be included in the experimental sample due to schools not being over-subscribed. **Randomization** = Student admission lotteries.

English language proficiency, special education status, and students in gifted/talented program. **Results** = Treatment had a  $0.066\sigma$  ( $0.035$ ) impact on math and a  $0.034\sigma$  ( $0.023$ ) impact on reading.

**Test score** = state math and reading tests. **Regression specification** = OLS regression controlling for student's age, gender, race/ethnicity, free lunch status, individualized education program status, pretreatment test scores, whether the student's primary home language is English, whether the household has only one adult, family income, mother's education, school fixed-effects, grade fixed-effects, and lottery year fixed-effects. We report average annual impacts. **Results** = Winning the lottery had a  $0.11\sigma$  ( $0.04$ ) impact on math test scores and a  $0.05\sigma$  ( $0.05$ ) impact on reading test scores.

*Continued*

**Table A3** Schools—cont'd**Study****Study design****Results**

Literacy learning of at-risk first-grade students in the reading recovery early intervention ([Schwartz, 2005](#)). N teachers = 37, N students = 148, grade = 1, location = 14 states. **Treatment groups** = treatment students participated in reading Recovery (RR). Control students did not.

**Treatment defined** = RR is an early intervention for low performing students. It consists of extensive professional development for the teachers and one-on-one 30 min daily lessons to accelerate the literacy learning of their students. Note that students initially assigned to the control group participated in RR after the completion of the experiment.

**Randomization** = two of lowest scoring students from each classroom were randomized into treatment or control.

**Treatment defined** = Treatment students took part in an after-school program designed to develop social and employment readiness, as well as boost academic performance. Students received \$1.25 for every hour they devoted to educational activities, as well as a significant reward if they enrolled in postsecondary education. Those students who graduated on-time received assistance in postsecondary placement. Treatment lasted a total of 750 h per year.

**Randomization** = Students were stratified by school and assigned at random to treatment.

**Test score** = Slosson oral reading test-Revised.

**Regression specification** = The effect size was calculated using posttest means.

**Results** = RR had a  $0.934\sigma$  ( $0.245$ ) impact on reading test scores.

Longer-Term Impacts of Mentoring, Educational Services, and Learning Incentives: Evidence from a Randomized Trial in the United States ([Rodríguez-Planas, 2012](#)). N recruitment sites = 7, N schools = 11, N students = 1,069, grades = 9–12, N years = 5. **Treatment groups** = treatment students received the Quantum opportunity program (QOP). Control students had access only to those opportunity programs available locally. Sample drawn from students whose 8th-grade GPA fell below the 67th percentile.

**Test score** = achievement tests developed by the national education Longitudinal study.

**Regression specification** = OLS regression controlling for gender, age, 8th-grade GPA, race/ethnicity, and school. We report average annual impacts.

**Results** = Treatment had a  $0.012\sigma$  ( $0.014$ ) impact on math test scores and a  $0.013\sigma$  ( $0.016$ ) impact on reading test scores.

Longitudinal Effects of Classwide Peer Tutoring (Greenwood et al., 1989). N schools = 6, N students = 416, grades = 1–4, N years = 4. **Treatment groups** = treatment group implemented Classwide peer tutoring (CWPT). Control group maintained normal curricula. Sample drawn from schools serving communities of low socioeconomic status.

Mastery learning and student teams: a factorial experiment in urban general mathematics (Slavin and Karweit, 1984). N schools = 16, N classrooms = 44, N students = 1,092, grade = 9, location = Philadelphia, PA.

**Treatment groups** = treatment classrooms implemented one of three curricula: a Mastery curriculum, a team-based curriculum, or both. The control classrooms utilized a focused instruction curriculum.

**Treatment defined** = at the start of each week, treatment students were assigned into tutor–tutee pairs, and these pairings were assigned to one of two teams. Tutees earned points for their team by completing tasks set by their tutors. Teachers determined the content to be tutored each week.

**Randomization** = four schools were assigned randomly to treatment; the remaining two schools were assigned to the control group.

**Treatment defined** = All classrooms used a standard course of instruction composed of 26 units. Students in the mastery condition were tested at the end of each unit to determine if they had achieved at least 80% mastery. Those who did not achieve mastery received remedial instruction, while those who did received enrichment instruction in the same unit. Students in the team-based condition were organized into four-member teams and quizzed each week. Team members' improvement in scores were summed to generate a team score, and the highest-scoring team was recognized each week. In the focused instruction condition, students worked individually and did not receive remedial instruction.

**Randomization** = Teachers were

**Test score** = the basic battery of the Metropolitan achievement test. **Regression specification** = posttest scores adjusted for pretest scores were used to calculate effect sizes. **Results** = Treatment had a  $0.106\sigma$  ( $0.206$ ) impact on math test scores and a  $0.162\sigma$  ( $0.209$ ) impact on reading test scores.

**Test Score** = Comprehensive Test of Basic Skills: mathematics computations and the concepts and applications subtests. **Regression specification** = average growth in test scores was used to calculate effect sizes.

**Results** = implementing both mastery and team conditions had a  $0.244\sigma$  ( $0.451$ ) impact on test scores, while implementing teams alone had a  $0.183\sigma$  ( $0.438$ ) impact, and implementing mastery alone had a  $0.015\sigma$  ( $0.403$ ) impact on test scores.

*Continued*

**Table A3** Schools—cont'd

Study	Study design	Results
<p>National Board Certification and Teacher Effectiveness: Evidence from a Random Assignment Experiment (Cantrell et al., 2008). N teachers = 198, N students <math>\approx</math> 3800, grades = 2–5, location = Los Angeles Unified school district.</p> <p><b>Treatment groups</b> = treatment classrooms were taught by teachers that had applied for certification at any point in time and control classrooms were taught by teachers that never applied.</p>	<p>stratified by school and assigned randomly to treatment.</p> <p><b>Treatment defined</b> = Treatment students were taught in classrooms where teachers had applied to the national Board for professional teaching standards (NBPTS) for certification. Note that to apply for the national Board certification teachers have to have a minimum of three years teaching experience. The researchers therefore restricted the control group to classrooms with teachers that had at least three years of experience. <b>Randomization</b> = Invitations were sent out to all elementary schools in the Los Angeles Unified school district and school participation was voluntary. Teachers in participating schools were matched to NBPTS records and grade 2–5 teachers that had ever applied for NBPTS certification were selected. The research team matched each of these teachers with another teacher in the same school, grade, and calendar track to serve as comparison. Principals were then asked to identify two classes that they would be willing to assign to either of these paired teachers. The researchers randomly assigned each pair of teachers to the classes that the principals chose for them. After classroom randomization, no further contact was made with the schools from the research team.</p>	<p><b>Test score</b> = California standards test: Math and language subtests. <b>Regression specification</b> = Student outcome regressions controlled for school-by-year-by-grade fixed effects, baseline math and reading scores interacted with grade, race/ethnicity, ever retained, Title I, eligible for free lunch, homeless, migrant, gifted and talented, special education, English language development, and the means of these variables among all students in the class.</p> <p><b>Results</b> = teachers who applied to certification had an impact of <math>-0.015\sigma</math> (0.071) on math test scores and an impact of <math>-0.019\sigma</math> (0.060) on reading test scores relative to teachers who never applied.</p>

Paying to learn: The effect of financial incentives on elementary school test scores (Bettinger, 2012). N students = 873, Grades = 3–6, location = Coshocton, OH.

**Treatment groups** = treatment students were incentivized by cash to pass tests. Control students were not incentivized.

Prevention and Remediation of Severe Reading Disabilities: Keeping the End in Mind (Torgesen et al., 1997). N schools = 13, N students = 180, grades = K-2, N years = 3.

**Treatment groups** = three treatment conditions: Condition one entailed phonological awareness training plus synthetic phonics instruction (PASP), condition two entailed implicit embedded phonics instruction (EP), and condition three entailed a regular classroom

**Treatment defined** = students received \$15 for each test on which they scored proficient or better. They received more for advanced or accelerated designation. Payment was given in the form of “Coshocton children’s bucks,” gift certificates redeemable at any store in Coshocton. **Randomization** = The unit of randomization was the grade level at each of four eligible elementary schools. Each year, eight of 16 eligible grade-school combinations were selected via lottery to receive financial incentives. First, the district randomly selected one grade per school. After these four drawings, Conschocton conducted a fifth drawing in which they chose four additional grade-school combinations from amongst the remaining possibilities.

**Treatment defined** = all treatment children received 80 min of supplemental individual instruction per week for 30 months. Children in the PASP condition received explicit instruction in how to sound-out words. Students in the EP group learned words through phonics games, contextual definitions, sentence construction, and reading exercises. The RCS group received tutoring in the skills and activities currently being taught in their curriculum. **Randomization** = students were randomly assigned to one of the four conditions.

**Test score** = Terra Nova or Ohio achievement standardized math and reading test. **Regression specification** = OLS regression controlling for grade, school, time fixed-effects, age, gender, race, free or reduced-price lunch status, pretest scores, and an indicator for outcome test taken. Results are pooled over three years. Standard errors are clustered at the grade-school level.

**Results** = Treatment had a  $0.1328\sigma$  (0.0485) impact on math test scores and a  $0.0103\sigma$  (0.0454) impact on reading test scores.

**Test score** = Woodcock Johnson Mastery Tests-Revised: Word Attack, word identification, and Passage comprehension subtests.

**Regression specification** = Average posttest scores were used to calculate effect sizes. Average annual effect across outcome measures is reported. **Results** = The PASP treatment had a  $0.286\sigma$  (0.103) impact on reading test scores. The EP treatment had a  $0.112\sigma$  (0.098) impact on

**Table A3** Schools—cont'd

Study	Study design	Results
<p>support group (RCS). The control group maintained normal curricula. Sample drawn entirely from children with low phonological language ability.</p>		<p>reading test scores. The RCS treatment had a <math>0.094\sigma</math> (<math>0.097</math>) impact on reading test scores.</p>
<p>Private school vouchers and student achievement: an evaluation of the Milwaukee parental choice program (Rouse, 1998). N students = 2,258, grades = K-8, location = Milwaukee, WI, N years = 4.</p> <p><b>Treatment groups</b> = treatment groups received a voucher to attend a private school. Control students did not receive a voucher. Sample composed entirely of students whose family income fell at least 1.75 times below the poverty line.</p>	<p><b>Treatment defined</b> = treatment students received vouchers to at least partially offset the cost of attendance at a nonsectarian private school of their choice.</p> <p><b>Randomization</b> = Students were stratified by the grade and school for which they were applying, then randomly assigned to treatment.</p>	<p><b>Test score</b> = The reading and math subtests of the Iowa test of basic skills. <b>Regression specification</b> = OLS regression controlling for school and grade applying, as well as gender and family income. Note we only report the results for one year after randomization due to the magnitude of attrition in later years. <b>Results</b> = Treatment had a <math>0.105\sigma</math> (<math>0.082</math>) impact on math test scores and a <math>0.049\sigma</math> (<math>0.075</math>) impact on reading test scores.</p>
<p>Putting Books in the Classroom Seems Necessary But Not Sufficient (McGill-Franzen et al., 1999). N schools = 6, N teachers = 18, N students = 456, grade = K, location = Large urban eastern</p>	<p><b>Treatment defined</b> = the professional development focused on techniques for encouraging children to pick up books and read them. The training covered topics such as physical design of the classroom, effective book displays, importance of reading aloud to children,</p>	<p><b>Test score</b> = Peabody Picture Vocabulary Test. <b>Regression specification</b> = Average gains from pre to posttest scores were used to calculate effect sizes. <b>Results</b> = The training and books treatment had a</p>

school district. **Treatment groups** = teachers in the first treatment group received training and books for their classroom. Teachers in the second treatment group did not receive training, but received books. Control teachers did not receive training or books.

Repeated Reading Intervention: Outcomes and Interactions with Readers' Skills and Classroom Instruction (Vadasy and Sanders, 2008). N schools = 13, N students = 162, grades = 2–3.

**Treatment groups** = Treatment students received the *Quick Reads* tutoring program. Control group received no tutoring. Sample drawn from students with demonstrated difficulty reading, as determined by pretest scores.

School Choice as a Latent Variable: Estimating the “Complier Average Causal Effect” of Vouchers in Charlotte (Cowen,

and small-group lessons using teacher made-materials. **Randomization** = eighteen kindergarten teachers, three each from six schools, were randomly assigned into three groups – (a) training and books, (b) no Training and books, (c) no Training and no books.

**Treatment defined** = the *Quick Reads* program builds reading fluency via repeated reading strategies, in which tutors first introduced a passage, and then students read it approximately three to four times. Tutoring sessions were conducted in pairs four days per week for 15 weeks. **Randomization** = Students were stratified by grade and school and placed at random into pairs. Pairings were then assigned at random to treatment.

**Treatment defined** = The vouchers offered grants of up to \$1700 annually to at least partially offset the cost of tuition at a private school. Treatment students could

$0.118\sigma$  (0.578) impact on reading test scores. The books treatment had a  $-0.465\sigma$  (0.585) on reading test scores.

**Test score** = Woodcock reading mastery tests-Revised: Word reading Accuracy subtest; test of word reading efficiency: Sight word subtest; Gary oral reading test 4: Rate and comprehension subtests.

**Regression specification** = For each outcome measure, effect sizes were calculated using the average growth between posttest and pretest scores. We report the average effect across all outcome measures.

**Results** = Treatment had a  $0.300\sigma$  (0.158) impact on reading test scores.

**Test score** = The Iowa Test of Basic Skills. **Regression specification** = OLS regressions controlling for

*Continued*

**Table A3** Schools—cont'd

Study	Study design	Results
<p>2008). N students = 1,143, grades = 2–8. Location = Charlotte, NC. <b>Treatment groups</b> = treatment students were offered a voucher to offset the cost of tuition at a private school. Control students applied but did not receive a voucher. Only low-income applicants were considered for the voucher program.</p> <p>School Choice in Dayton, Ohio after Two Years: An Evaluation of the Parents Advancing Choice in Education Scholarship Program (West et al., 2001). N students = 515, grades = 2–9, location = Dayton, OH.</p> <p><b>Treatment groups</b> = students from the treatment group received scholarships to help with the cost of private schools. Control group comprises of students who lost lottery for scholarships. Analysis sample consists of only those students who were in public school at the time of randomization. Vouchers were also offered to students already in private school.</p>	<p>attend the private school of their choice. <b>Randomization</b> = random lottery.</p> <p><b>Treatment defined</b> = Parents advancing choice in education offered low income parents scholarships to help defray the costs of sending their children to private schools in Dayton, Ohio. <b>Randomization</b> = Random lottery.</p>	<p>family income, mother's education, mother's race, whether both parents lived at home, and student's gender. <b>Results</b> = Being offered a voucher had a <math>0.237\sigma</math> (0.131) impact on math test scores and a <math>0.292\sigma</math> (0.134) impact on reading scores.</p> <p><b>Test score</b> = Iowa Test of Basic Skills. <b>Regression specification</b> = OLS regression controlling for pretest math and reading scores. We report annual impacts. <b>Results</b> = Treatment had a <math>0.054\sigma</math> (0.112) impact on math test scores and a <math>0.078\sigma</math> (0.112) impact on reading test scores.</p>

School Choice in New York City After Three Years: An Evaluation of the School Choice Scholarships Program (Mayer et al., 2002). N families = 1,960, grades = 1–5, location = New York city.

**Treatment groups** = treatment families were offered a scholarship funded by the school choice scholarships foundation (SCSF). Control families were not offered a scholarship. To be eligible for the scholarship, students had to be eligible for free or reduced-price lunch.

Summer school effects in a randomized field trial (Zvoch and Stevens, 2012). N students = 93, grades = K-1.

**Treatment groups** = Treatment students were invited to participate in a summer literacy program. Control students were not.

**Treatment defined** = Treatment families received \$1400 annually from the SCSF for at least three years. The scholarship was designed to at least partially offset the cost of private school attendance. Families could select the private school of their choice. **Randomization** = eligible applicants were stratified by whether their school's test scores were above or below the city-wide median. Eighty-five percent of the treatment group was selected at random from those applicants whose schools were below the median. The remaining 15% was selected at random from those applicants whose schools were above the median.

**Treatment defined** = The literacy program was a five-week program that lasted for 3.5 h a day, four days a week. In the program, students received classroom instruction on fundamental literacy topics, were assigned homework, completed in-class work packets, and practiced literacy skills in small groups with students of a similar skill level.

**Randomization** = In the years preceding the intervention, all students below certain cutoff scores on the Nonsense word fluency test or test of oral reading fluency were invited to participate in summer school. In 2010, all students

**Test score** = The reading and mathematics subtests of the Iowa Test of Basic Skills.

**Regression Specification** = OLS regression controlling for pretest scores and whether the student came from a public school whose test scores were below the median. We report the average annual impact over three years. **Results** = Treatment had a  $0.020\sigma$  ( $0.033$ ) impact on math test scores and a  $0.003\sigma$  ( $0.033$ ) impact on reading test scores.

**Test score** = dynamic indicators of basic early literacy: Nonsense word fluency subtest for kindergarten students and the test of oral reading fluency for the first grade students.

**Regression specification** = OLS regressions. The ITT results reported by the researchers do not contain any covariates as controls. However, they do note that in models that contained student characteristic variables, the treatment effects remained

*Continued*

**Table A3** Schools—cont'd**Study****Study design****Results**

Teacher Behavior and Pupil Performance: Reconsideration of the Mediation of Pygmalion Effects (Alpert, 1975). N schools = 13, N teachers = 17, N classrooms = 17, N students = 352, grade = 2, location = New York city.  
**Treatment groups** = treatment teachers were asked to increase the frequency of certain behaviors. Control teachers received no such intervention. Sample drawn exclusively from Catholic schools.

Teacher incentives and student achievement: evidence from New York city public schools

that fell below the cutoff scores were invited and not included in the analysis. The district then established upper bounds so that approximately 50 kindergartners and 50 first graders fell in the range between the cutoff scores and the upper bound scores. Students that fell in this range of scores were considered the experimental sample and randomized into treatment or control.

**Treatment defined** = over a period of 11 weeks, treatment teachers were asked to increase target behaviors. These target behaviors included more reading group time, maximizing “best” reading time — those periods when the teacher reported feeling most motivated to teach, covering more materials in their reading group, including fewer pupils in their reading group, and utilizing more good verbal behaviors (praising students’ reading, support, reinforcement, praising students’ behavior, encouraging questions, etc.) — and positive reinforcement of student success. **Randomization** = schools were randomly assigned to treatment.

**Treatment defined** = Treatment involved giving schools financial incentives based on whether they met the annual

qualitatively similar.  
**Results** = Treatment had a  $0.691\sigma$  (0.280) impact on reading test scores.

**Test score** = The vocabulary and reading comprehension subtests of the Gates-MacGintie Reading Test.

**Regression**

**Specification** = For each outcome measure, effect sizes were calculated using the average growth between posttest and pretest scores. We report the average effect across all outcome measures.

**Results** = treatment had a  $0.072\sigma$  (0.557) impact on reading test scores.

**Test score** = State math and reading test scores. **Regression specification** = Regressions

(Fryer, 2014b). N schools = 396, N students = 185,612, grades = K-12, location = New York city. **Treatment groups** = Treatment schools received financial incentives. Control schools did not.

Teacher Pay for Performance: Experimental Evidence from the Project on Incentives in Teaching (Springer et al., 2010). N teachers = 296, N students = 23,784, grades = 5–8. **Treatment groups** = treatment teachers participated in the project on incentives in teaching (POINT) and control teachers did not.

performance target set by the Department of Education. Schools were free to distribute money among teachers at their own discretion.

**Randomization** = Schools were randomly assigned based on average proficiency ratings, poverty rates and student demographic characteristics. Final experimental sample consisted of 233 treatment schools and 163 control schools.

**Treatment Defined** = POINT was open to middle school mathematics teachers. POINT allowed for treatment teachers to receive an incentive for sufficiently high value added. All treatment and control teachers received a stipend of \$750.

**Randomization** = Schools were stratified into 10 groups based on student scores in prior years. Randomization was done within strata. Clusters of teachers were then assigned to treatment or control status. Clusters were defined based on course groups. Assignments to treatment and control were permanent for the duration of the project.

include test scores from previous years, demographic characteristics and school level controls. We report the annual impact. **Results** = Treatment had a  $-0.030\sigma$  (0.019) impact on math test scores and a  $-0.018\sigma$  (0.021) impact on reading test scores.

**Test Score** = Tennessee Comprehensive Assessment Program. **Regression Specification** = Linear

models controlling for pretest scores, race/ethnicity, gender, English Language Learner status, special education status, free/reduced-price lunch status, number of days of suspension and unexcused absences, teacher's value-added from the year before the experiment, and the average pretest score of students in a teacher's class. The models also include block fixed effects and cluster random effects. We report the average annual impact across three years.

**Results** = Treatment had an

*Continued*

**Table A3** Schools—cont'd**Study****Study design****Results**

Teacher Study Group: Impact of the Professional Development Model on Reading Instruction and Student Outcomes in First Grade Classrooms ([Gersten et al., 2010](#)). N recruitment sites = 3, N schools = 19, N teachers = 84, N students = 575, Grade = 1.

**Treatment Groups** = Treatment teachers took part in a teacher study group (TSG) intervention. Control teachers utilized the district's normal professional development program. Sample drawn from schools already utilizing the *Reading First* program.

Team pay for performance: experimental evidence from the round rock pilot project on team incentives ([Springer et al., 2012](#)). N schools = 9, N teacher

**Treatment Defined** = Treatment entailed a professional development program designed to integrate research-based comprehension and vocabulary instruction into the classroom. Teachers in treatment schools read and discussed instructional methods in small groups and designed curricula to incorporate these strategies. Treatment occurred over sixteen 75-min sessions held twice a month from October to June.  
**Randomization** = Schools were assigned at random to treatment. Seven students from each classroom were selected at random for the student sample.

**Treatment Defined** = Performance awards were distributed to teams of teachers based on their collective contribution to student test score gains in the four core subjects. Team performance

impact of  $0.017\sigma$  (0.023) on math test scores and an impact of  $0.003\sigma$  (0.012) on reading test scores.

**Test Score** = Woodcock Diagnostic Reading Battery: Oral Vocabulary, Reading Vocabulary, Passage Comprehension, Letter Word Identification, and Word Attack subtests; Dynamic Indicators of Basic Early Literacy Skills: Oral Reading Fluency subtests. **Regression Specification** = Two-level hierarchical linear model (student, teacher) controlling for pretest scores, as well as teacher and school fixed effects. We report the average effect across all outcome measures. **Results** = Treatment had a  $0.225\sigma$  (0.205) impact on reading test scores.

**Test Score** = Texas Assessment of Knowledge and Skills; Stanford Achievement Test. **Regression Specification** = Two-level hierarchical linear

teams = 159, N students = 17,383, Grades = 6–8, Location = Round Rock Independent School District in Texas. **Treatment Groups** = Treatment teachers received monetary awards based on their students' performance and control teachers did not.

The (Surprising) Efficacy of Academic and Behavioral Intervention with Disadvantaged Youth: Results from a Randomized Experiment in Chicago (Cook et al., 2014). N students = 106, Grades = 9–10, Location = South Side of Chicago. **Treatment Groups** = Students in treatment one participated in "Becoming a Man" (BAM). Students in treatment two participated in BAM and received tutoring. Control students continued business as usual. The sample consisted entirely of male students.

was based on a value-added measure of student performance or standardized test scores. Teams were predefined by the district and were organized such that each team had at least one teacher for each core subject. Each team typically oversaw the learning experience of 100–140 students. **Randomization** = In each year of the 2-year study, teams were randomized to either the awards intervention or control condition using block-randomization design. Blocks were defined by grades within school.

**Treatment Defined** = BAM included in-school programming that exposes youth to prosocial adults, and provides them with social-cognitive skill training that follows the principles of cognitive behavioral therapy. Students in treatment two also received high-dosage, small-group tutoring. **Randomization** = From the student population, 106 high-risk males were chosen to participate based on an "academic risk index." These students were stratified by grade and then randomly assigned to the three groups.

model (student, team) controlling for pretreatment test scores and demographics. We report the average impact across years and outcomes.

**Results** = Treatment had an impact of  $0.000\sigma$  (0.020) on math test scores and an impact of  $-0.002\sigma$  (0.017) on reading test scores.

**Test Score** = EXPLORE and PLAN tests, developed by ACT Inc. **Regression Specification** = An OLS regression controlling for age, grade, prior math and reading test scores, Individualized Education Plan status, previous year GPA, absences, suspensions and disciplinary incidents, and free lunch eligibility. **Results** = Treatment one had a  $0.611\sigma$  (0.227) impact on math test scores and a  $-0.071\sigma$  (0.319) impact on reading test scores. Treatment two had a  $0.425\sigma$  (0.226) impact on math test scores and a  $-0.043\sigma$  (0.262) impact on reading test scores.

*Continued*

**Table A3** Schools—cont'd**Study****Study design****Results**

<p>The effect of school choice on participants: evidence from randomized lotteries (<a href="#">Cullen et al., 2006</a>). N districts = 1, N schools = 19, N students = 14434, Grade = 8, Location = Chicago, N years = 2. <b>Treatment</b>  <b>Groups</b> = Treatment students received an offer of admission to a school of their choice. Control students did not receive admission. Sample composed of oversubscribed high schools that determined admission via random lottery.</p> <p>The Effective Instruction of Comprehension: Results and Description of the Kamehameha Early Education Program (<a href="#">Tharp, 1982</a>). N schools = 2, N classrooms = 8, Grade = 1, Location = HI. <b>Treatment</b></p>	<p><b>Treatment Defined</b> = This study uses randomized lotteries that determine high school admission in Chicago Public Schools to investigate the impact of school choice. The authors exploit the fact that Chicago's public school students can apply to gain access to public schools outside of their neighborhood school (this is known as an open enrollment system).  <b>Randomization</b> = Oversubscribed schools stratify applicants by gender and race, and offer admission via a random lottery.</p> <p><b>Treatment Defined</b> = The KEEP intervention is a small-group program designed to boost reading comprehension among at-risk students. The program utilizes face-to-face student–teacher interaction to teach comprehension instruction, as well as sight vocabulary and</p>	<p><b>Test Score</b> = Reading subtest of the Test of Academic Proficiency. <b>Regression Specification</b> = OLS regression controlling for race, pretest scores, age, free lunch eligibility, special education, bilingual education, living with biological parent, attending assigned 8th grade school, census tract characteristics (fraction Black, fraction Hispanic, poverty rate, fraction high school graduates, fraction homeowners, fraction not in labor force, crime index, fraction of high school students attending private schools), and lottery fixed-effects.  <b>Results</b> = Winning a lottery to any school had a <math>-0.038\sigma</math> (0.015) impact on reading test scores.  <b>Test Score</b> = The Gates-MacGintie Reading Test and the Metropolitan Achievement Test. <b>Regression Specification</b> = Effect sizes were calculated using the average posttest scores. We</p>
--	--	--

**Groups** = Treatment classrooms implemented the Kamehameha Early Education Program (KEEP) intervention. Control classrooms received no such intervention. Sample drawn entirely from semirural public schools.

The effectiveness of extended day programs: evidence from a randomized field experiment in the Netherlands (Meyer and Klaveren, 2013). N schools = 7, N students = 188, Grades = 5–7, Location = Netherlands.

**Treatment Groups** = Treatment students received an offer to participate in an extended school day program. The control group received no such offer.

The Effectiveness of Secondary Math Teachers from Teach for America and the Teaching Fellows Programs (Clark et al., 2013). N states = 11, N districts = 15, N schools = 82, N teachers = 287, N students = 12,699, Grades = 6–12. **Treatment Groups** = Two treatment groups: Students in treatment one were taught by

analytic phonics. The intervention took the place of normal class time.

**Randomization** = Students were randomly assigned to treatment.

**Treatment Defined** = The extended day program consisted of an additional 2 h of language instruction, 2 h of math instruction, and 1 h of excursions per week. The intervention was conducted at one of the participating schools, and classes were composed of 10 students.

**Randomization** = Students were randomly assigned to treatment or control.

**Treatment Defined** = TFA and Teaching Fellows programs take a distinctive approach to addressing the need for high-quality teachers of hard-to-staff subjects in high-poverty schools. TFA and the Teaching Fellows programs have highly selective admissions criteria designed to admit only applicants who have demonstrated a high level of achievement in academics or other endeavors and who possess characteristics that the programs

report the average effect across all tests. **Results** = Treatment had a  $0.300\sigma$  (0.141) impact on reading test scores.

**Test Score** = Standardized tests typically used in Dutch elementary schools.

**Regression Specification** = OLS regressions controlling for math pretest scores, gender, minority status, parents' education, family structure, and class size. **Results** = Assignment to treatment had a  $0.087\sigma$  (0.067) impact on math test scores and a  $0.005\sigma$  (0.081) impact on reading test scores.

**Test Score** = For middle school students, authors obtained scores on state-required assessments. For high school students, authors administered end-of-course math assessments developed by the Northwest Evaluation Association. **Regression Specification** = OLS regressions that controlled for

**Table A3** Schools—cont'd

Study	Study design	Results
<p>teachers from the Teach for America (TFA) program. Students in treatment two were taught by teachers from the New Teacher Project Teaching Fellows program. Control students were taught by teachers who did not enter teaching through either of these programs.</p> <p>The Effectiveness of Team-Accelerated Instruction on High Achievers in Mathematics (<a href="#">Karper and Melnick, 1993</a>). N classrooms = 12, N students = 247, Grades = 3–5, Location = Hershey, PA.</p> <p><b>Treatment Groups</b> = Treatment classrooms implemented the Team-Accelerated Instruction (TAI) math program. Control classrooms continued with their normal curricula.</p>	<p>view as being associated with effective teaching. <b>Randomization</b> = In each participating school, authors identified “classroom matches” — two or more classes covering the same middle or high school math course at the same level, with at least one class taught by a teacher from the program being studied (TFA or Teaching Fellows) and at least one class taught by another teacher, referred to as a comparison teacher, who did not enter teaching through a highly selective alternative route. Students were randomly assigned to these classes.</p> <p><b>Treatment Defined</b> = The TAI program creates a competitive classroom environment, in which teams of students earn points by completing common goals. Students progress through the program’s subject matter as rapidly as they are able. <b>Randomization</b> = Four classrooms per grade were included in the study, and two classrooms within each grade were assigned to the treatment condition. Students were stratified by grade and randomly assigned to treatment or control classrooms.</p>	<p>students’ pretest scores, baseline characteristics, and classroom match indicators. <b>Results</b> = Assignment to TFA teachers had a <math>0.07\sigma</math> (0.02) impact on math test scores. Assignment to Teaching Fellows had a <math>0.00\sigma</math> (0.02) impact on math test scores.</p> <p><b>Test Score</b> = Iowa Test of Basic Skills: math concepts and math computation subtests.</p> <p><b>Regression Specification</b> = Effect sizes were calculated using the average growth between posttest and pretest scores.</p> <p><b>Results</b> = Treatment had a <math>-0.037\sigma</math> (0.227) impact on math scores.</p>

The Effects of A One-Year Staff Development Program on the Achievement Test Scores of Fourth-Grade Students (Cole, 1992). N schools = 1, N teachers = 12, N students = 268, Grade = 4, Location = MS. **Treatment Groups** = Treatment teachers received the Mississippi Teacher Assessment Instrument staff development program. Control teachers continued business as usual.

The Effects of “Brain Gym” as a General Education Intervention: Improving Academic Performance and Behaviors (Nussbaum, 2010). N students = 364, Grades = 2–6, Location = East TX. **Treatment Groups** = Treatment students were assigned to classrooms that used the “Brain Gym” curriculum. Control students were assigned to classrooms that continued with their normal curricula.

**Treatment Defined** = Treatment teachers underwent a comprehensive staff development training program using Mississippi Teacher Assessment Instrument modules as training materials. The 14 Mississippi Teacher Assessment Instrument teacher (pedagogical) behavior competencies include topics such as planning instruction to achieve selected objectives, organizing instruction to take into account individual differences among learners, and obtaining and using information about the needs and progress of individual learners.

**Randomization** = Teachers were randomly assigned to the treatment or control group.

**Treatment Defined** = Brain Gym is a movement based program designed to promote whole-brain learning. It is derived from the fundamental premise that learning occurs as humans receive sensory stimuli and initiate movement.

**Randomization** = Students were randomly assigned to classrooms and then classrooms were randomly assigned to treatment and control groups.

**Test Score** = Math and reading scores from the Stanford Achievement Test.

### Regression

**Specification** = Effect sizes were calculated using average posttest score adjusted for pretest scores. **Results** = Treatment had a  $0.508\sigma$  ( $0.586$ ) impact on math test scores and a  $0.566\sigma$  ( $0.589$ ) impact on reading test scores.

**Test Score** = The math and reading subtests of the Texas Assessment of Knowledge and Skills. **Regression**

**Specification** = Effect sizes were calculated from the growth between pre and posttest means. **Results** = Brain Gym had a  $0.130\sigma$  ( $0.105$ ) impact on math test scores and a  $0.188\sigma$  ( $0.106$ ) impact on reading test scores.

*Continued*

**Table A3** Schools—cont'd  
**Study**

	<b>Study design</b>	<b>Results</b>
<p>The effects of peer-assisted literacy strategies for first-grade readers with and without additional mini-skills lessons (<a href="#">Mathes and Babyak, 2001</a>). N schools = 5, N teachers = 30, N students = 130, Grade = 1.</p> <p><b>Treatment Groups</b> = Two treatment groups: one incorporated Peer-Assisted Learning Strategies (PALS) into their curricula, the other incorporated PALS and small-group mini-lessons. Control group maintained their normal curricula.</p>	<p><b>Treatment Defined</b> = During PALS sessions, higher-ability students were paired with lower-ability students as determined by the teacher. These pairings would earn points by completing assigned tasks. These 35-min sessions occurred three times per week for 14 weeks. Students in the mini-lesson group were placed into groups of three and received 15–20 min of fluency instruction from their teachers three times weekly for six weeks. All interventions took the place of normal classtime instruction.</p> <p><b>Randomization</b> = Schools were stratified by demographic similarity. Researchers determined the number of teachers to recruit from each type of school to create a representative stratified sample. Volunteer teachers were randomly assigned to the three groups.</p>	<p><b>Test Score</b> = The Woodcock Reading Mastery Tests-Revised. <b>Regression Specification</b> = For each outcome measure, effect sizes were calculated using the average growth between posttest and pretest scores. We report the average effect across all outcome measures.</p> <p><b>Results</b> = The PALS treatment had a <math>0.779\sigma</math> (0.465) impact on reading test scores. The PALS and mini-lessons treatment had a <math>0.854\sigma</math> (0.499) impact on reading test scores.</p>
<p>The Effects of Structured One-on-One Tutoring in Sight Word Recognition of First-Grade Students At Risk for Reading Failure (<a href="#">Mayfield, 2000</a>). N students = 60, Grade = 1, Location = LA. <b>Treatment Groups</b> = Treatment students received 15 min per day of one-on-one tutoring and control students were read to aloud in small groups for 15 min per day.</p>	<p><b>Treatment Defined</b> = Tutoring was done as a part of the Edmark Reading Program, where America Reads volunteers were trained for 2 h either individually or in small groups by the researcher.</p> <p><b>Randomization</b> = Principals and teachers selected a sample of students from the bottom 20–30% of first grade readers. Students in the selected sample were then randomly assigned to treatment or control groups.</p>	<p><b>Test Score</b> = Woodcock Reading Mastery Tests-Revised: Word Identification and Passage Comprehension subtests. <b>Regression Specification</b> = Effect sizes were calculated using average posttest scores. <b>Results</b> = Treatment had a <math>0.346\sigma</math> (0.184) impact on reading test scores.</p>

The Effects of Theoretically Different Instruction and Student Characteristics on the Skills of Struggling Readers (Mathes et al., 2005). N schools = 6, N students = 298, Grade = 1. **Treatment Groups** = Treatment one students participated in the Proactive Reading program. Treatment two students participated in the Responsive Reading program. Control students did not participate in any such program.

The Efficacy of an Early Literacy Tutoring Program Implemented by College Students (Allor and McCathren, 2004). N students = 137, Grade = 1. **Treatment Groups** = Treatment group received

**Treatment Defined** = All interventions took place outside of normal class for 40 min per day, five days per week, in groups of three. Proactive Reading systematically built reading skills to develop fluency. As the intervention progressed, words gradually became more complicated and texts became more difficult. Responsive Reading had teachers tailor daily lessons to student needs. Teachers offered explicit instruction in reading and gradually let students become more independent as the intervention progressed. All schools in the study used an enhanced curriculum that built upon the district's normal curriculum by offering assessment measures to help teachers identify if and how students were struggling with reading.

**Randomization** = The sample was drawn from students at risk of developing persistent reading difficulty, as determined by the pretest. Eligible students were stratified by school and then assigned at random to one of the three groups.

**Treatment Defined** = College students administering the tutoring to treatment group students received three 1-h group training sessions and additional assistance on site. Each student was tutored on average 2 to 3 times per week for 15–20 min per session. **Randomization** =

**Test Score** = The Woodcock –Johnson reading battery.

**Regression Specification** = For each outcome measure, effect sizes were calculated using the posttest means. We report the average effect across all outcome measures.

**Results** = The *Proactive Reading* treatment had a  $-0.067\sigma$  (0.157) impact on math test scores and a  $0.283\sigma$  (0.158) impact on reading test scores. The *Responsive Reading* treatment had a  $-0.133\sigma$  (0.156) impact on math test scores and a  $0.250\sigma$  (0.156) impact on reading test scores.

**Test Score** = Woodcock Johnson-Revised: Word Identification, Word Attack, and Passage Comprehension subtests; Test of Word Reading Efficiency: Real Word and Nonword subtest; Dynamic

*Continued*

**Table A3** Schools—cont'd**Study****Study design****Results**

<p>tutoring and control group received no tutoring.</p> <p>The Evaluation of Charter School Impacts: Final Report (Gleason et al., 2010). N schools = 36, N students = 2330. <b>Treatment Groups</b> = The treatment group were students that won lotteries to attend charter schools. The control group were students that lost those same lotteries.</p>	<p>At-risk students were identified by low test scores and teacher recommendations. Eligible students were then randomly selected for either the treatment or control group.</p> <p><b>Treatment Defined</b> = This study investigated the impact of charter schools on students' achievement. The researchers focused on charter schools with fourth to seventh grade entry grades and that were at least two years old. Thirty-six charter schools (from multiple states) were eligible and willing to participate with 2005–2006 or 2006–2007 entry cohorts. <b>Randomization</b> = Student admission lotteries. To be considered in the experimental sample, students had to apply to one of the 36 charter schools during an experimental lottery and give consent to participate in the study.</p>	<p>Indicators of Basic Early Literacy Skills: Phoneme-Segmentation Fluency and Nonsense Word Fluency subtests. <b>Regression Specification</b> = Effect sizes were calculated using average posttest scores. We report the average effects size across cohorts and outcome measures. <b>Results</b> = Treatment had a <math>0.422\sigma</math> (0.222) impact on reading test scores.</p> <p><b>Test Score</b> = State math and reading tests. <b>Regression Specification</b> = OLS regressions controlling for pretest reading and math achievement, disciplinary measures, student demographics, family characteristics, school enrollment, and application history. We report the average annual impact of winning a lottery to one of these 36 charter schools. <b>Results</b> = Admission to a charter school had a <math>-0.03\sigma</math> (0.03) impact on math test scores and a <math>0.04\sigma</math> (0.03) impact on reading test scores.</p>
--	--	--

The Evaluation of Enhanced Academic Instruction in After-School Programs: Final Report (Black et al., 2009). N schools = 27, N students = 1,218, Grades =

### 2–5. Treatment

**Groups** = Treatment group incorporated enhanced academic instruction into their after-school programs. Control group maintained their regular after-school programs. All schools had preexisting after-school programs.

The Impact of Elementary Mathematics Coaches on Student Achievement (Campbell and Malkus, 2011). N districts = 5, N schools = 36, N classrooms = 1,169, N

**Treatment Defined** = The enhanced program entailed 45 min of structured instruction at the start of the two- to 3-h after-school program; this formal instruction took the place of passive academic support like homework help or tutoring. The goal of this enhanced instruction was the development of new skills, as opposed to the completion of assignments. All after-school programs were offered four days per week. Schools had their choice of implementing either the reading or math programs based on student needs, but were limited to one.

**Randomization** = To be eligible for the study, students had to perform at most two years below grade-level in reading or math. Eligible students who applied to the program were then stratified by after-school center and by grade and then assigned at random to treatment. All participants were either already enrolled in after-school programs or referred to such programs based on their poor performance.

**Treatment defined** = Coaches attended five mathematics courses in numbers and operations, geometry and measurement, algebra and functions, and probability and statistics before entering their designated school. Coaches also attended a leadership

**Test Score** = Stanford Achievement Test: Mathematics and Reading subtests. **Regression**

**Specification** = The standardized difference of means was calculated for each outcome measure. The means were adjusted for pretest scores, gender, race/ethnicity, free or reduced-price lunch status, age, whether the student is from a single-parent household, whether the student was overage for their grade, and mother's education level. We report the average effect across outcome measures and cohorts.

**Results** = Treatment had a  $0.09\sigma$  (0.04) impact on math test scores and a  $-0.04\sigma$  (0.05) impact on reading test scores.

**Test Score** = The Standards of Learning assessment (the standardized state assessment of Virginia). **Regression**

**Specification** = Three-level hierarchical linear model

*Continued*

**Table A3** Schools—cont'd

Study	Study design	Results
<p>students = 24,759, Grades = 3–5, Location = VA, N years = 3. <b>Treatment</b>  <b>Groups</b> = Treatment schools received a mathematics coach to work with their teachers. Control schools received no such intervention.</p> <p>The Impact of Indiana's System of Interim Assessments on Mathematics and Reading Achievement (<a href="#">Konstantopoulos et al., 2013</a>). N schools = 57, N students ≈ 20,000, Grades = K-8, Location = IN. <b>Treatment Groups</b> = Treatment schools participated in Indiana's Diagnostic Assessment Tools. Control schools continued as usual.</p>	<p>training course one year after entering their school. Coaches worked with teachers to design both curricula and assessments as well as facilitate classtime instruction. <b>Randomization</b> = Each district sorted its schools into groups of three based on their demographic composition and history of performance on mathematics assessments. Two schools from each group of three were selected randomly for treatment.</p> <p><b>Treatment Defined</b> = In 2008, the Indiana Department of Education introduced the Diagnostic Assessment Tools. This program consisted of two commercial products, mCLASS (grades K-2) and Acuity (grades 3–8). With mCLASS, teachers are provided with detailed diagnostic measures of their K-2 students in literacy and numeracy. Acuity provides teachers with multiple-choice online assessments in reading and mathematics for Grades 3–8. The assessments are approximately 30 min long, typically completed in groups in</p>	<p>(student, class, school) controlling for age, gender, English-proficiency status, special education status, free/reduced-lunch status, minority status, whether the teacher had a masters degree, teacher's experience at the school, Title I eligibility, school size, and indicators for past academic performance. We report the average effect across all cohorts and grades. <b>Results</b> = Treatment had a <math>0.049\sigma</math> (<math>0.090</math>) impact on math test scores.</p> <p><b>Test Score</b> = For grades 3–8, the mathematics and reading ISTEP<sup>+</sup> (Indiana's state test). In grades K-2, the math and reading portions of Terra Nova. <b>Regression Specification</b> = A two-level hierarchical linear model (student, school) controlling for gender, age, race, socioeconomic status, special education status, limited English proficiency status, and school-level percentages of</p>

The Impact of Two Professional Development Interventions on Early Reading Instruction and Achievement (Garet et al., 2008). N states = 4, N districts = 6, N schools = 90, N teachers = 270, N students  $\approx$  5,000, Grade = 2.

**Treatment Groups** = Two treatment groups: Teachers in treatment one participated in a reading content-focused professional development program. Teachers in treatment two participated in the same professional development and additionally received in-school

class, and aligned to the state standards. When a school adopts these products, their teachers are provided with training on how to effectively use them.

**Randomization** = The pool of eligible schools were placed into 4 blocks based on locales. From these blocks, 70 schools were randomly drawn. Eleven of these schools were dropped due to previous use of products from the vendors in the study or a school closure. The remaining 59 schools were randomized into an intervention group or a control group in an unbalanced manner (35 treatment and 24 control).

**Treatment Defined** = Teachers in both treatment groups participated in a teacher institute series that began in the summer and continued through the school year. Throughout the course of the year, treatment teachers attended eight seminar days that each consisted of 6 h of instruction for a total of 48 h of professional development. On top of this, teachers in treatment two also received approximately 60 h of in-school coaching.

**Randomization** = Schools were randomly assigned to treatment one, treatment two, or control such that there were equal numbers of schools assigned to each group in a given district. In five of the districts, schools were grouped into blocks

females, minorities, lower socioeconomic status, and limited English proficiency students. **Results** = Treatment had an impact of  $0.127\sigma$  ( $0.069$ ) on math test scores and  $0.078\sigma$  ( $0.050$ ) impact on reading test scores, respectively.

**Test Score** = The reading state test scores used in a given district. **Regression**

**Specification** = Regressions controlling for school level pretest scores, student-level gender, age, race/ethnicity, and a separate poverty measure provided by each district.

**Results** = The professional development treatment had a  $0.08\sigma$  ( $0.08$ ) impact on reading test scores. The professional development plus coaching treatment had a  $0.03\sigma$  ( $0.09$ ) impact on reading test scores.

**Table A3** Schools—cont'd

Study	Study design	Results
<p>coaching. Control teachers did not receive any professional development or coaching.</p> <p>The Influence of Massive Rewards on Reading Achievement in Potential Urban School Dropouts (<a href="#">Clark and Walberg, 1968</a>). N classrooms = 9, N students = 110, Ages = 10–13.</p> <p><b>Treatment Groups</b> = The treatment group increased the amount of verbal praise rendered to each student. The control group maintained its normal level of verbal praise.</p>	<p>of similar characteristics (percentage of minority students or geographic region, depending on the district) and then one-third of each block was randomly assigned to each treatment group. The remaining district was just randomly split into thirds.</p> <p><b>Treatment Defined</b> = Students received rewards during remedial reading sessions in the form of verbal praise; students reported the number of times they received such praise daily. Teachers in the treatment group were asked to at least double the amount of verbal praise rendered to each student.</p> <p><b>Randomization</b> = To be eligible for the study, students had to score one to four years behind grade level on nationally-standardized achievement tests—ranking them as potential dropouts. These students were assigned randomly to one of nine after-school remedial reading programs. Five of these classrooms were assigned at random to treatment.</p>	<p><b>Test Score</b> = Science Research Associates Reading Test, Intermediate Form.</p> <p><b>Regression Specification</b> = Effect sizes were calculated using posttest means of the outcome measure. <b>Results</b> = Treatment had a <math>0.588\sigma</math> (<math>0.685</math>) impact on reading test scores.</p>
<p>The Potential of Urban Boarding Schools for the Poor: Evidence from SEED (<a href="#">Curto and Fryer, 2014</a>). N students = 221.</p> <p><b>Treatment Groups</b> = Treatment students received admission to a SEED school;</p>	<p><b>Treatment Defined</b> = SEED schools are five-day-a-week urban boarding schools that have an extended school day, provide extensive after-school tutoring, utilize data-driven curricula, and maintain a culture of high expectations. The middle schools focus on developing basic math</p>	<p><b>Test Score</b> = The DC CAS test.</p> <p><b>Regression Specification</b> = OLS regression controlling for student pretest scores, gender, free lunch eligibility, special education status, and English language learner status.</p>

control students applied but did not receive admission to a SEED school. All students included in the sample were black.

The Prevention, Identification, and Cognitive Determinants of Math Difficulty (Fuchs et al., 2005). N schools = 10, N classrooms = 41, N students = 127,

Grade = 1. **Treatment**

**Groups** = Treatment group received math tutoring in addition to normal class time. The control group continued with their normal curricula. Sample drawn from students deemed at risk of developing mathematics difficulty.

The Reading Connection: A Leadership Initiative Designed to Change the Delivery of Educational Services to At-Risk Children (Compton, 1992). N students = 483, Grade = 1, Location = Kalamazoo, MI.

and reading skills, while high schools utilize a college-preparatory curriculum that requires students to take the SAT or ACT, as well as apply to at least five colleges. This study utilizes the fact that when a SEED school is oversubscribed, it determines admission via a random lottery. **Randomization** = Admission to an oversubscribed SEED school is determined by random lottery stratified by gender.

**Treatment Defined** = Students selected for treatment received extra math tutoring immediately following their normal mathematics instruction. Tutoring occurred in small groups three times per week for 16 weeks in addition to regular class time. The purpose of this tutoring was to curb mathematics difficulty before it developed. **Randomization** = Students were stratified by classroom and assigned at random for treatment.

**Treatment Defined** = The *Reading Connection* program is an early intervention designed to curb reading difficulty before it leads to persistent failure in school. The intervention entailed individual tutoring for 30 min per day, five days per week, for 12–16

**Results** = Winning the lottery had a  $0.218\sigma$  (0.082) impact on math test scores and a  $0.201\sigma$  (0.086) impact on reading test scores.

**Test Score** = Woodcock –Johnson III: Applied Problems and Computation subtests. **Regression Specification** = Average growth in test scores was used to calculate effect sizes. The average effect across both subtests is reported. **Results** = Treatment had a  $0.300\sigma$  (0.179) impact on math test scores.

**Test Score** = The Iowa Test of Basic Skills. **Regression Specification** = Effect sizes were calculated using posttest means. **Results** = Treatment had a  $0.216\sigma$  (0.092) impact on reading test scores.

**Table A3** Schools—cont'd

Study	Study design	Results
<p><b>Treatment Groups</b> = Treatment group implemented the <i>Reading Connection</i> program. Control group maintained traditional remedial reading services. Sample drawn from students already enrolled in small-group remedial reading instruction.</p> <p>Transfer Incentives for High-Performing Teachers: Final Results from a Multisite Randomized Experiment (<a href="#">Glazerman et al., 2013</a>). N states = 7, N districts = 10, N students <math>\approx</math> 7000. <b>Treatment Groups</b> = Treatment students were taught by high-quality teachers who filled vacancies through a transfer incentive program. Teaching vacancies in control schools were filled as usual, without incentives.</p>	<p>weeks. The goal of this tutoring was to both build reading fluency and develop self-monitoring skills so that students can assess and resolve their own difficulties in the future. <b>Randomization</b> = Students were assigned at random to treatment.</p> <p><b>Treatment Defined</b> = Talent Transfer Initiative allows for the highest-performing teachers in each district to receive a pay raise to move into schools serving the most disadvantaged students. Teachers ranking in roughly the top 20% within their subject and grade were offered \$20,000 if they transferred and remained in a set of designated schools. <b>Randomization</b> = Researchers first identified low-achieving schools that had a vacancy within a teaching team. Whenever possible, schools were matched within district based on the grade and subject of the vacancy and student demographics. Within these matched blocks, schools were randomly assigned to either treatment or control status.</p>	<p><b>Test Score</b> = Math and reading state tests. <b>Regression Specification</b> = Regression model controlling for pretest scores, race/ethnicity, gender, English Language Learner status, special education status, free/reduced-price lunch status, over-age for grade status, an indicator of whether the student belonged to a study team that had at least one retention-stipend teacher, grade dummies, and block dummies. <b>Results</b> = Treatment had a <math>0.120\sigma</math> (<math>0.070</math>) impact on math test scores and a <math>0.058\sigma</math> (<math>0.050</math>) impact on reading test scores.</p>

Using Knowledge of Children's Mathematics Thinking in Classroom Teaching: An Experimental Study (Carpenter et al., 1989). N schools = 24, N teachers = 40, N students ≈ 480, Grade = 1, Location = Madison, WI.

### **Treatment**

**Groups** = Treatment group attended a 4-week workshop; control group attended two 2-h workshops.

When Less May Be More: A 2 Year Longitudinal Evaluation of a Volunteer Tutoring Program Requiring Minimal Training (Baker et al., 2000). N students = 84, Grades = 1 –2.

### **Treatment**

**Groups** = Students in the treatment group received one-on-one tutoring and the control group received normal classroom instruction.

**Treatment Defined** = The treatment workshop helped teachers learn how children develop addition/subtraction skills, focusing on cognitive processes applied to different word problems. Teachers then developed strategies to build math skills utilizing these processes. The workshop met 5 h a day, four days a week for the first four weeks of the teachers' summer vacation. Control workshops focused on nonroutine problem solving. **Randomization** = Stratified by school, teachers were randomly assigned to treatment or control. Six male and six female students were randomly selected from each class for the analysis. In two instances, there were less than 12 total students in a class, in which case the entire class was used.

**Treatment Defined** = Tutors received a 1 –2 h training session either at the beginning of the school year or anytime throughout. Treatment group students attend tutoring sessions for 30 min twice a week during the school year. Students were tested at the beginning of first grade, the end of first grade, and the end of 2nd grade. **Randomization** = Eligible students were referred by teachers. They were students with poor reading skills and little academic experiences with adults at

**Test Score** = Iowa Test of Basic Skills-Level 7: Computation and Mathematics Problems subtests. **Regression**

**Specification** = Effect sizes were calculated for the two subtests using means adjusted for pretest scores. We report the average effect size on these two tests. **Results** = Treatment had a  $0.396\sigma$  ( $0.319$ ) impact on math test scores.

**Test Score** = Woodcock Reading Mastery Tests-Revised: Word Identification, Word Comprehension and Passage Comprehension subtests. **Regression**

**Specification** = Effect sizes were calculated using posttest means adjusted for pretest scores. We report the average annual impact across all outcome measures.

*Continued*

**Table A3** Schools—cont'd**Study****Study design****Results**

When Schools Stay Open Late:  
 The National Evaluation of the  
 21st Century Community  
 Learning Centers Program  
 (James-Burdumy et al., 2005). N  
 districts = 12, N centers = 26,  
 N students = 2308. **Treatment**  
**Groups** = Treatment students  
 were able to enroll in 21st  
 Century after-school centers.  
 Control students were unable to  
 enroll in these centers for two  
 years after randomization.

home or otherwise. Eligible students were randomly assigned to treatment and control groups through a process called Rapid Letter Naming.

**Treatment Defined** = The 21st Century Community Learning Centers ran an after-school program for treatment students. Centers offered homework sessions, academic activities, and enrichment activities.

**Randomization** = Random assignment was conducted at the center level. Applicants to each of the centers were randomized into treatment and control. For seven sites, random assignment took place at the beginning of the 2000–2001 school year and for the other five sites, random assignment took place at the beginning of the 2001–2002 school year.

**Results** = Treatment had a  $0.184\sigma$  (0.156) impact on reading test scores.

**Test Score** = Stanford Achievement Test.

**Regression Specification** = Regression model controlling for student characteristics.

**Results** = Treatment had a  $-0.021\sigma$  (0.046) impact on math test scores and a  $0.008\sigma$  (0.046) impact on reading test scores.

**Table A4** Curriculum Study

	Study Design	Results
<p>A Mixed-Method Multilevel Randomized Evaluation of the Implementation and Impact of an Audio-Assisted Reading Program for Struggling Readers (<a href="#">Lesnick, 2006</a>). N districts = 2, N schools = 9, N classrooms = 59, N students = 233, Grades = 3 and 5. <b>Treatment Groups</b> = Treatment classrooms implemented the <i>New Heights</i> reading intervention. Control classrooms continued with their normal curricula. Sample drawn from students who were at least nine months below grade level in reading, as determined by the pretest.</p>	<p><b>Treatment Defined</b> = The <i>New Heights</i> program builds reading fluency via repeated reading strategies with audio assistance. Students select their own book from the interventions' reading list, the teacher gives a brief introduction to introduce new vocabulary or skills, and then the student reads through the text once. Students then elect to re-read the book with or without the audio assistance, complete a worksheet connected to the text, or conference with the teacher. The intervention lasted approximately 20–30 min per day, five days per week for 18 weeks.</p> <p><b>Randomization</b> = Students were sorted into blocks by school, grade, classroom, and gender. Half the students in each block (up to a maximum of six) were assigned randomly to treatment. The remainder from each block were assigned to the control group.</p>	<p><b>Test Score</b> = Dynamic Indicator of Basic Early Literacy Skills; Test of Word Reading Efficiency. <b>Regression Specification</b> = MANCOVA analysis controlling for pretest scores, gender, race/ethnicity, free/reduced lunch status, and English-language-learner status. Effect sizes were averaged across outcome measures. <b>Results</b> = Treatment had a <math>-0.028\sigma</math> (0.108) impact on reading test scores.</p>
<p>A Multisite Cluster Randomized Field Trial of Open Court Reading (<a href="#">Borman et al., 2008</a>). N schools = 6, N classrooms = 57, N students = 1,099, Grades = 1–5. <b>Treatment Groups</b> = Treatment classrooms implemented the <i>Open Court Reading</i> (OCR) curriculum.</p>	<p><b>Treatment Defined</b> = The OCR curriculum provides textbooks, workbooks, decodable texts, and anthologies to develop reading fluency. The core components of the curriculum include: preparing to read—which builds phonemic awareness, sound and letter familiarity, phonics, fluency, and word knowledge; reading and</p>	<p><b>Test Score</b> = TerraNova Comprehensive Test of Basic Skills V: Reading Comprehension and Vocabulary subtests. <b>Regression Specification</b> = Effect sizes were calculated using average posttest scores. <b>Results</b> = Treatment had a <math>0.187\sigma</math> (0.288) impact on reading test scores.</p>

*Continued*

**Table A4** Curriculum—cont'd

Study	Study Design	Results
Control classrooms continued with their normal curricula.	<p>responding—which builds textual-thinking skills, vocabulary, reading proficiency, as well as comprehension, inquiry, and investigation strategies; and language arts — which develops writing skills, spelling, grammar usage and mechanics, vocabulary, penmanship, and listening. <b>Randomization</b> = Classrooms were placed into blocks by school and grade. Within each block, classrooms were randomly assigned to treatment.</p>	
<p>A Multisite Cluster Randomized Trial of the Effects of CompassLearning Odyssey Math on the Math Achievement of Selected Grade 4 Students in the Mid-Atlantic Region: Final Report (<a href="#">Wijekumar et al., 2009</a>). N schools = 32, N teachers = 122, N students = 2,854, Grade = 4, Locations = DE, NJ, and PA. <b>Treatment Groups</b> = Treatment teachers received the CompassLearning Odyssey Math program. Control teachers continued with business as usual.</p>	<p><b>Treatment Defined</b> = Odyssey Math is a computer-based math curriculum developed by CompassLearning Inc. to improve math learning for K-12 students. The software consists of a web accessed series of learning activities, assessments, and math tools. CompassLearning professional development trainers presented the learning activities, math tools, and assessments as available options to intervention teachers during the summer professional development session. Five days of Odyssey Math professional development were purchased for each treatment teacher, consisting of two large group presentations and three in class coaching sessions.</p> <p><b>Randomization</b> = A volunteer sample of teachers and their classrooms were randomly assigned to treatment and control conditions within schools.</p>	<p><b>Test Score</b> = TerraNova Basic Battery: Math subtest. <b>Regression Specification</b> = Multilevel hierarchical linear model (student, teacher, school) controlling for pretest scores.</p> <p><b>Results</b> = Treatment had a <math>0.02\sigma</math> (0.03) impact on math test scores.</p>

A Randomized Experimental Evaluation of the Impact of Accelerated Reader/Reading Renaissance Implementation on Reading Achievement in Grades 3 to 6 (Nunney et al., 2006). N teachers = 44, N students = 1,023, Grades = 3–6. **Treatment**

**Groups** = Treatment teachers incorporated both the *Accelerated Reader* program and the *Reading Renaissance* program. Control teachers maintained their normal curricula.

A Randomized Field Trial of the Fast ForWord Language Computer-Based Training Program (Borman et al., 2009). N schools = 8, N students = 415, Grades = 2 and 7, Location = Baltimore, MD.

**Treatment Groups** = Treatment students received supplemental Fast ForWord program instruction. Control students did not.

**Treatment Defined** = The *Accelerated Reader* program is a software-based curriculum, in which students select a book of their choice and complete reading comprehension quizzes. The program identifies weaknesses in reading comprehension and suggests texts to redress these difficulties. The *Reading Renaissance* program is a professional development program, which suggests teachers incorporate 30–60 min of reading time in class. It also trains teachers in the appropriate use of the *Accelerated Reader* software. **Randomization** = Teachers were randomly assigned to pairs within grade level, and then one teacher from each pairing was assigned randomly to treatment.

**Treatment Defined** = Fast ForWord Language is an adaptive computer program for language instruction. It is designed to build oral language comprehension skills and other critical skills necessary to become a better reader.

**Randomization** = Students were deemed eligible for the intervention if they scored below the 50th percentile on the Total Reading outcome for the district administered Comprehensive Test of Basic Skills, Fifth Edition. Eligible students were stratified by school and grade level and assigned randomly to treatment or control.

**Test Score** = The STAR Reading test.

**Regression Specification** = Effect size was calculated using the average growth between pre and posttest scores. We report the average effect across all grades.

**Results** = Treatment had a  $0.182\sigma$  ( $0.302$ ) impact on reading test scores.

**Test Score** = Comprehensive Test of Basic Skills V: Language and Reading Comprehension subtests. **Regression**

**Specification** = OLS regression controlling for pretest scores, teacher rating of students' abilities, demographic controls, and school fixed effects.

**Results** = Treatment had a  $0.088\sigma$  ( $0.168$ ) impact on reading test scores.

*Continued*

**Table A4** Curriculum—cont'd

Study	Study Design	Results
<p>A Study on the Effects of Houghton Mifflin Harcourt's Journeys Program: Year 1 Final Report (<a href="#">Resendez and Azin, 2012</a>). N schools = 6, N teachers = 44, N students = 1,046, Grades = K-2. <b>Treatment Groups</b> = Treatment classrooms implemented the Journeys program. Control classrooms continued with their normal curricula.</p>	<p><b>Treatment Defined</b> = The Journeys program is a comprehensive reading and language arts curriculum. The program includes reading, writing, and grammar exercises segmented into thematic units. Weekly lessons focused on one unit for five weeks, creating continuity among lesson content. <b>Randomization</b> = Teachers were stratified by school and randomly assigned to treatment.</p>	<p><b>Test Score</b> = Iowa Test of Basic Skills. <b>Regression Specification</b> = Three-level hierarchical linear model (changes over time, student, teacher). <b>Results</b> = Treatment had a <math>0.175\sigma</math> (0.049) impact on reading test scores.</p>
<p>Action Research: Implementing <i>Connecting Math Concepts</i> (<a href="#">Snider and Crawford, 1996</a>). N classrooms = 2, N students = 46, Grade = 4. <b>Treatment Groups</b> = Treatment classrooms implemented the <i>Connecting Math Concepts</i> (CMC) curriculum. Control classrooms continued with their normal curricula.</p>	<p><b>Treatment Defined</b> = The CMC curriculum develops multiple math skills within each lesson, stressing the relationship between skills. Students thus learn to develop complex mathematical procedures with simple math skills. <b>Randomization</b> = Students were assigned at random to the treatment or control classrooms.</p>	<p><b>Test Score</b> = The National Achievement Test. <b>Regression Specification</b> = Effect size was calculated using the average growth between posttest and pretest scores. <b>Results</b> = Treatment had a <math>0.258\sigma</math> (0.296) impact on math test scores.</p>
<p>An Efficacy Study on Scott Foresman's <i>Reading Street</i> Program: Year One Report (<a href="#">Wilkerson et al., 2006</a>). N districts = 3, N schools = 5, N teachers = 48, N students = 944, Grades = 1–3. <b>Treatment Groups</b> = Teachers in the treatment group included the <i>Reading Street</i> Program in their curriculum. The control group did not alter their curriculum. No teachers had previously used <i>Reading Street</i> materials.</p>	<p><b>Treatment Defined</b> = The <i>Reading Street</i> Program provides reading materials and a curriculum designed to develop critical reading skills: phonemic awareness, phonics, vocabulary, comprehension, and fluency. Teachers administer the program for at least 90 min in class. The program also recommends that students needing additional help receive up to 30 min of small-group tutoring, and those students who continue to struggle after such intervention receive individualized attention outside of normal class time. <b>Randomization</b> = Teachers were stratified by school and grade and then assigned randomly to treatment or control.</p>	<p><b>Test Score</b> = Gates-MacGintie Reading Test, Fourth Edition; Dynamic Indicators of Basic Early Literacy: Oral Reading Fluency subtest. <b>Regression Specification</b> = A two-level hierarchical linear model (student, school) controlling for gender, ethnicity, special education status, grade, and school. <b>Results</b> = Treatment had a <math>-0.095\sigma</math> (0.103) impact on reading test scores.</p>

An Evaluation of the Effects of Paired Learning in a Mathematics Computer-Assisted-Instruction Program (Turner, 1985). N teachers = 4, N classrooms = 12, N students = 275, Grades = 3–4, Location = Goodyear, AZ.

**Treatment Groups** = Classrooms in the first treatment group had students take part in computer-assisted-instruction individually. Classrooms in the second treatment group had students take part in computer-assisted-instruction in pairs. Control classrooms continued with their normal curricula.

An Experimental Study of the Effects of the Accelerated Reader Program and a Teacher Directed Program on Reading Comprehension and Vocabulary of Fourth and Fifth Grade Students (Knox, 1996). N schools = 1, N students = 77, Grades = 4–5. **Treatment**

**Groups** = All students received the same reading list. The treatment group reviewed their books with both the researcher and the Accelerated Reader computer program. The control group reviewed their books with their teachers. No student had previously used the Accelerated Reader program.

**Treatment Defined** = The computer-assisted instruction software presents the material, evaluates the student response, and progressively adjusts the material based on these responses. Instruction lasted for 15 minutes, three days per week in lieu of normal class time. Students working in pairs helped each other find the correct answer. Pairings were rotated so that, by the end of the experiment, all students in a class had worked together at some point. Students in the individual treatment worked alone and directed questions toward the teacher.

**Randomization** = Classrooms were stratified by teacher and assigned at random to one of the three conditions.

**Treatment Defined** = The Accelerated Reader program is designed to turn reading into a game. Upon completion of a book, a computer program tests students on reading comprehension and awards points based on these tests. The program also offers feedback based on these test results. Students select their own books. Rewards were awarded when a student reached a specific point threshold. **Randomization** = Students were stratified by grade and then paired across classrooms according to their pretest score. One student from each pair was assigned randomly to the treatment group.

**Test Score** = Comprehensive Test of Basic Skills: Mathematics subtests. **Regression Specification** = Average growth in test scores was used to calculate effect sizes.

**Results** = The individual treatment had a  $0.278\sigma$  (0.711) impact on math scores. The paired treatment had a  $0.395\sigma$  (0.714) impact on math scores.

**Test Score** = Stanford Achievement Test: Vocabulary and Reading Comprehension subtests. **Regression Specification** = Effect sizes were calculated for each outcome measure using means adjusted for pretest scores.

The researchers report the results by grade and subtests. We report the weighted average of the effects for subtests and grades. **Results** = Treatment had a  $-0.115\sigma$  (0.324) impact on reading test scores.

*Continued*

**Table A4** Curriculum—cont'd

Study	Study Design	Results
<p>Comparative Effectiveness of Scott Foresman Science: A Report of a Randomized Experiment in Five School Districts (<a href="#">Miller and Jaciw, 2007</a>). N districts = 5, N teachers = 92, N students = 2,638, Grades = 3–5. <b>Treatment</b>  <b>Groups</b> = Treatment teachers used the Scott Foresman Science curriculum. Control teachers continued with their normal curricula.</p>	<p><b>Treatment Defined</b> = Scott Foresman Science is a year-long science curriculum. The curriculum provides materials for both students and teachers aimed at developing independent investigative skills. In addition to science instruction, the curriculum features Leveled Reader. These are student readers designed to provide the teacher with an easy way to differentiate instruction and provide reading support at, below, and above grade level. Treatment teachers were provided a one-half day workshop with the materials for the curriculum.</p> <p><b>Randomization</b> = Volunteer teachers within each district were assigned to treatment or control by coin toss.</p>	<p><b>Test Score</b> = Northwest Evaluation Association Test: Reading achievement subtest. <b>Regression Specification</b> = The effect size was calculated using posttest means adjusted for pretest scores. <b>Results</b> = The impact of treatment on reading tests was <math>0.05\sigma</math> (0.04).</p>
<p>Computer Assisted Instruction as an Enhancer of Remediation (<a href="#">Hotard and Cortez, 1983</a>). N students = 190, Grades = 3–6. <b>Treatment</b>  <b>Groups</b> = Treatment students incorporated 10 min daily of computer-assisted-instruction (CAI) into their curricula. Control classrooms continued with their normal curricula.</p>	<p><b>Treatment Defined</b> = For six months, treatment students received 10 min of CAI instruction for mathematics daily in addition to their normal math lab instruction. Each lesson had students solve a variety of problems based on the material currently being taught in class. The software automatically adjusted the difficulty of its problems based on student performance. <b>Randomization</b> = Students were matched by their pretest scores and then one student from each pairing was assigned randomly to treatment.</p>	<p><b>Test Score</b> = Comprehensive Test of Basic Skills. <b>Regression Specification</b> = Effect size was calculated using the average growth between pre and posttest scores. <b>Results</b> = Treatment had a <math>0.193\sigma</math> (0.145) impact on math test scores.</p>

Computer-Assisted Instruction to Prevent Early Reading Difficulties in Students at Risk for Dyslexia: Outcomes from Two Instructional Approaches ([Torgesen et al., 2010](#)). N students = 112, Grade = 1.

**Treatment Groups** = Two treatment conditions: condition one implemented the *Read, Write, and Type* program; condition two implemented the *Lindamood Phoneme Sequencing Program for Reading, Spelling, and Speech*. Control students received no such intervention.

Costs, Effects, and Utility of Microcomputer Assisted Instruction ([Fletcher et al., 1990](#)). N schools = 1, N classrooms = 4, N students = 60, Grades = 3 and 5, Location = Saskatchewan, Canada. **Treatment Groups** = Treatment classrooms utilized microcomputer assisted mathematics instruction. Control classroom continued with their normal curricula.

**Treatment Defined** = The interventions were implemented over four, 50-min sessions per week from October through May. The first half of each lesson was devoted to direct reading instruction from teachers, and the remaining time was devoted to practicing these skills on the computer. Students in condition one received lessons in alphabetic reading skills, while students in condition two received explicit instruction in phonemic awareness. **Randomization** = Students were stratified by school and then randomly assigned to one of the three groups.

**Treatment Defined** = Third-grade treatment students used the *Milliken Math Sequences* software to practice mathematics skills introduced in class for an average of 12 min per day, five days per week. Fifth-grade treatment students utilized the software for mathematics practice, drilling up to 15 min, four days per week. All computerized mathematics practice took the place of normal class time. **Randomization** = One classroom from each grade was designated the treatment classroom, while the other served as control. Students were assigned at random to the treatment or control classrooms in their grade.

**Test Score** = Comprehensive Test of Phonological Processing; Woodcock Reading Mastery Test-Revised: Word Identification, Word Attack, and Passage Comprehension subtests; Test of Word Reading Efficiency: Word Efficiency and Phonemic Decoding subtests.

**Regression Specification** = Effect sizes were calculated using posttest means. We report the average impact across all outcome measures. **Results** = The *Read, Write, and Type* program had a  $0.459\sigma$  (0.238) impact on reading test scores. The *Lindamood Phoneme Sequencing Program for Reading, Spelling, and Speech* program had a  $0.702\sigma$  (0.240) impact on reading test scores.

**Test Score** = Canadian Test of Basic Skills.

**Regression Specification** = Effect sizes were calculated using posttest scores adjusted for pretest scores. We report the average effect across grades. **Results** = Treatment had a  $0.421\sigma$  (0.322) impact on math test scores.

*Continued*

**Table A4** Curriculum—cont'd

Study	Study Design	Results
<p>Does Rainbow Repeated Reading Add Value to an Intensive Intervention Program for Low-progress Readers? An Experimental Evaluation (Wheldall, 2000). N sites = 2, N students = 40, Grades = 2–7, Location = Australia. <b>Treatment Groups</b> = All students were participating in an intensive literacy intervention. Treatment students received a supplemental literacy program, the <i>Rainbow Reading Program</i>. Control students continued with their normal curricula.</p>	<p><b>Treatment Defined</b> = All students included in the study were classified as low-progress readers and were attending a literacy program focused on repeated reading. The treatment students supplemented this program with the <i>Rainbow Reading Program</i> – a repeated reading program where students read along with an audio tape.  <b>Randomization</b> = Within each site, students were ranked by baseline reading accuracy and were paired with a student of comparable reading ability. One student of each pair was randomly assigned to treatment and the other was assigned to control.</p>	<p><b>Test Score</b> = The Burt Word Reading Test. <b>Regression Specification</b> = Effect sizes were calculated using the average growth between pre and posttest scores. <b>Results</b> = Treatment had a <math>0.018\sigma</math> (0.105) impact on reading test scores.</p>
<p>Effectiveness of Reading and Mathematics Software Products: Findings From Two Student Cohorts (Campuzano et al., 2009). N districts = 27, N schools = 105, N teachers = 347, N students = 9,392, Grades = 1, 4, and 6. <b>Treatment Groups</b> = The eight treatment groups all used different software products. Teachers in the control group continued using their normal curricula.</p>	<p><b>Treatment Defined</b> = Treatment classrooms were given the following software products: First grade reading: Destination Reading, Waterford Early Reading Program, Headsprout and Plato Focus; 4th grade reading: LeapTrack and Academy of Reading; 6th grade math: Larson PreAlgebra and Achieve Now.  <b>Randomization</b> = Districts volunteered and identified schools. Then teachers volunteered and were assigned to treatment and control.</p>	<p><b>Test Score</b> = Each district's nationally normed tests. If a district did not administer a standardized test, they used Stanford Achievement Test. Only administered tests to one randomly selected treatment and control classroom per school. <b>Regression Specification</b> = Two-level hierarchical model (student, classroom) controlling for pretest scores, age, gender, teachers' years of experience, education level, and school fixed effects. <b>Results</b> = The impacts for each software product are as follows: Destination Reading = <math>0.091\sigma</math> (<math>0.082</math>); Headsprout = <math>0.014\sigma</math> (<math>0.052</math>); Plato Focus = <math>0.024\sigma</math> (<math>0.066</math>); Waterford Early Reading Program = <math>0.020\sigma</math> (<math>0.068</math>); Academy of Reading = <math>-0.008\sigma</math> (<math>0.050</math>); LeapTrack = <math>0.094\sigma</math> (<math>0.036</math>); Larson PreAlgebra = <math>0.113\sigma</math> (<math>0.076</math>); Achieve Now = <math>-0.028\sigma</math> (<math>0.069</math>).</p>

Effectiveness of Selected Supplemental Reading Comprehension Interventions: Impacts on a First Cohort of Fifth-Grade Students (James-Burdumy et al., 2008). N districts = 10, N schools = 89, N teachers = 268, N students = 6,350, Grade = 5. **Treatment**

**Groups** = The four treatment groups all used different software products. Control schools continued with their normal curricula.

Effects of Health-Related Physical Education on Academic Achievement: Project SPARK (Sallis et al., 1999). N schools = 7, N students = 759, Grades = 4–5, Location = Southern CA.

**Treatment Groups** = In the first treatment, specialists conduct physical education programs. In the second treatment, a specialist trains classroom teachers to conduct physical education classes. The control group continues with its current physical education program. All participating schools did not employ physical education specialists prior to the study.

Effects of Targeted Intervention on Early Literacy Skills of At-Risk Students

**Treatment Defined** = The four reading comprehension curricula were Project CRISS, ReadAbout, Read for Real, and Reading for Knowledge. Districts involved in the study were required to have at least 12 Title I schools and schools must not already have been using any of the four curricula. **Randomization** = Schools were randomly assigned to one of the four treatment groups or control group within each district. When possible, schools in a district were blocked into groups with similar pretest scores and then randomization occurred within each block. When blocks were not possible, a Chromy selection procedure was implemented.

**Treatment Defined** = Physical education classes were administered to 4th graders three days a week throughout the school year; classes were designed to promote independent physical activity via weekly fitness goals and family involvement. The classes continued through 5th grade.

**Randomization** = Schools were stratified by the percentage of ethnic minority students and then randomly assigned into one of the three groups.

**Treatment Defined** = The authors developed their own intervention

**Test Score** = Group Reading Assessment and Diagnostic Evaluation. **Regression Specification** = OLS regression model that controls for student pretest scores, English language learner status, race/ethnicity, teacher race, school urbanicity, and district fixed effects.

**Results** = Project CRISS had a  $-0.04\sigma$  (0.04) impact on reading test scores. ReadAbout had a  $-0.07\sigma$  (0.05) impact on reading test scores. Read for Real had a  $-0.06\sigma$  (0.04) impact on reading test scores. Reading for Knowledge had a  $-0.11\sigma$  (0.04) impact on reading test scores.

**Test Score** = Metropolitan Achievement Tests: Language and Reading subtests.

**Regression Specification** = Effect sizes were calculated using the average growth between pre and posttest scores. We report the average effect across subtests. We only report results from the first cohort because there appears to be an error in the results reported for the second cohort in the published article.

**Results** = The specialist treatment had a  $-0.006\sigma$  (0.456) impact on math scores and a  $0.101\sigma$  (0.458) impact on reading scores. The trained teacher treatment had a  $0.000\sigma$  (0.456) impact on math scores and a  $0.103\sigma$  (0.458) impact on reading scores.

**Test Score** = Woodcock Reading Mastery Test-Revised: Word Identification,

**Table A4** Curriculum—cont'd

Study	Study Design	Results
<p>(Wang and Algozzine, 2008). N schools = 6, N students = 139, Grade = 1. <b>Treatment Groups</b> = Treatment students replaced their remedial reading instruction with a targeted literacy intervention. Control students continued with their normal remedial instruction. Sample drawn from students at risk of persistent reading failure, as determined by the pretest.</p>	<p>targeted at the following skills: phonemic awareness, letter-sound correspondence, reading phonetically, fluency building, and sight-word practice. Instruction lasted approximately 10–15 min daily in class. <b>Randomization</b> = Randomly selected two schools to serve as control schools.</p>	<p>Word Attack, and Passage Comprehension subtests; Dynamic Indicators of Basic Early Literacy Skills: Phoneme Segmentation Fluency and Nonsense Word Fluency subtests. <b>Regression Specification</b> = For each outcome measure, effect sizes were calculated using the average growth between posttest and pretest scores. We report the average effect across all outcome measures. <b>Results</b> = Treatment had a <math>0.303\sigma</math> (0.873) impact on reading test scores.</p>
<p>Efficacy of Collaborative Strategic Reading with Middle School Students (Vaughn et al., 2011). N recruitment sites = 2, N schools = 6, N teachers = 17, N classrooms = 61, N students = 866, Grades = 7–8. <b>Treatment Groups</b> = Treatment classrooms implemented the <i>Collaborative Strategic Reading (CSR)</i> intervention. Control classrooms did not implement such an intervention.</p>	<p><b>Treatment defined</b> = The CSR intervention teaches students to reflect upon their comprehension of the text. The intervention follows a set strategy: first, students reviewed the topic of the text prior to reading; second, students read the text, noting where they had difficulty; third, students assessed how they were able to overcome those difficulties; fourth, students worked in small groups to discuss the text and any unresolved difficulties. The intervention took place in class for 50 min per day, twice weekly, over 18 weeks. <b>Randomization</b> = Students were randomly assigned to classes, and classes were stratified by teacher and randomly assigned to treatment.</p>	<p><b>Test Score</b> = The Gates-MacGintie Reading Test. <b>Regression Specification</b> = Effect size was calculated using the average growth between pre and posttest scores. <b>Results</b> = Treatment had a <math>0.073\sigma</math> (0.072) effect on reading test scores.</p>

Empirical Evaluation of *Read Naturally* Effects (Christ and Davie, 2009).

N districts = 4, N schools = 6, N students = 109, Grade = 3.

**Treatment Groups** = The treatment group used *Read Naturally* software for 30 min daily in class. The control group engaged in nonreading activities during the same time. No school had previously used the *Read Naturally* program.

Evaluation Research on the Effectiveness of *Fluency Formula*: Final Report (Sivin-Kachala and Bialo, 2005). N districts = 2, N classrooms = 12, N students = 128, Grade = 2.

**Treatment Groups** = The treatment group implemented the *Fluency Formula* reading curriculum. The control group maintained their normal classroom curricula. No classrooms had access to *Fluency Formula* materials beforehand.

**Treatment Defined** = The *Read Naturally* program utilizes computer software to build reading fluency via the following three strategies: repeated reading, reading with a model, and progress monitoring with feedback. The program also builds vocabulary via reading-for-meaning strategies. **Randomization** = To be eligible for the study, students had to fall below the 40th percentile on both a standardized test of oral fluency (either the Dynamic Indicators of Basic Early Literacy Skills or the AIMSweb Test of Early Literacy) and a measure of reading comprehension (Measures of Academic Progress) administered at the end of 2nd grade. Eligible students who agreed to participate were stratified by classroom and assigned at random to the treatment group.

**Treatment Defined** = *Fluency Formula* builds oral fluency by focusing on the following units: partner reading, choral reading, expressive reading, reading theater, repeated reading, and expert reading. In-class instruction lasts for at least 15 min per day, five days per week; 15-min take-home assignments are also given once per week. Students requiring additional instruction receive small-group tutoring in class for at least 15 min.

**Randomization** = Numerous teachers from each district volunteered for the study; researchers matched pairs of classrooms with similar reading ability, ethnic composition, and teacher characteristics. One classroom from each pair was randomly assigned to treatment.

**Test Score** = Dynamic Indicators of Basic Early Literacy Skills; Test of Word Reading Efficiency; Gray Oral Reading Tests IV: Reading Fluency and Accuracy measures; Woodcock Reading Mastery Test-Revised: Word Identification subtest. **Regression Specification** = Effect sizes were calculated for each outcome measure using posttest means adjusted for pretest covariates. We report the average effect across all outcome measures. **Results** = Treatment had a  $0.248\sigma$  (0.197) impact on reading test scores.

**Test Score** = Woodcock-Johnson III: Passage Comprehension subtest.

**Regression Specification** = The effect size was calculated using the growth between pre and posttest means.

**Results** = Treatment had a  $-0.271\sigma$  (0.178) impact on reading test scores.

**Table A4** Curriculum—cont'd

Study	Study Design	Results
<p>Fostering the Development of Vocabulary Knowledge and Reading Comprehension Through Contextually-Based Multiple Meaning Vocabulary Instruction (<a href="#">Nelson and Stage, 2007</a>). N classrooms = 16, N students = 283, Grades = 3 and 5. <b>Treatment Groups</b> = Treatment classrooms altered their curricula to include contextual vocabulary instruction. Control classrooms continued with their normal curricula.</p>	<p><b>Treatment Defined</b> = Treatment classrooms incorporated into their curricula contextually-based multiple meaning vocabulary instruction, which taught students to derive the meanings of words from a given context. The treatment was designed to boost vocabulary and reading comprehension. Instruction lasted approximately 20–30 min per day during each school day. <b>Randomization</b> = Classrooms were stratified by grade and then randomly assigned for treatment.</p>	<p><b>Test Score</b> = Gates-MacGintie Reading Tests IV: Vocabulary and Reading Comprehension subtests. <b>Regression Specification</b> = Effect sizes were calculated for each outcome measure using the growth between pre and posttest means. We report the average effect across subtests. <b>Results</b> = The treatment had a <math>0.205\sigma</math> (0.502) impact on reading test scores.</p>
<p>Impact of Thinking Reader Software Program on Grade 6 Reading Vocabulary, Comprehension, Strategies, and Motivation (<a href="#">Drummond et al., 2011</a>). N schools = 32, N teachers = 92, N students = 2,140, Grade = 6, Locations = CT, MA, and RI. <b>Treatment Groups</b> = Treatment teachers received three Thinking Reader digital novels to read with their students and participated in professional development to learn how to use the software. Control teachers used the schools' regular curricula.</p>	<p><b>Treatment Defined</b> = Thinking Reader is software in which students read novels on computers and respond to prompts from animated coaches. Treatment teachers received two, 6-h training sessions in the fall and three individual coaching sessions throughout the school year (<math>\approx 8</math> h). <b>Randomization</b> = Within each school, sixth grade reading/ELA teachers were randomly assigned to treatment or control.</p>	<p><b>Test Score</b> = Gates-MacGintie Reading Tests: Vocabulary and Reading Comprehension subtests. <b>Regression Specification</b> = Multilevel hierarchical linear model (student, teacher, school) controlling for students' pretest scores, English language learner status, special education status, teachers' education and years of experience, school poverty, and school size. <b>Results</b> = Treatment had a <math>-0.005\sigma</math> (0.053) impact on reading test scores.</p>
<p>Improving Reading Comprehension and Social Studies Knowledge in Middle School (<a href="#">Vaughn et al., 2013</a>). N schools = 2, N teachers = 5, N classes = 27, N students = 419, Grade = 8.</p>	<p><b>Treatment Defined</b> = PACT is a program designed to improve text comprehension and content learning. The model has 5 key components: 1) A comprehension canopy that contains a motivational springboard and an overarching issue or</p>	<p><b>Test Score</b> = Gates-MacGinitie: Reading Comprehension subtest. <b>Regression Specification</b> = A multilevel, multiple-group structural equation model was used to create latent estimates for student outcomes. The effect size was calculated</p>

**Treatment Groups** = Treatment classrooms adopted a Promoting Acceleration of Comprehension and Content Through Text (PACT) model. Control classrooms continued with their normal curricula.

Improving Reading Fluency and Comprehension in Elementary Students Using Read Naturally (Arvans, 2009). N schools = 1, N students = 82, Grades = 2–4.

**Treatment Groups** = Treatment students utilized the *Read Naturally* software. Control students received no such intervention. Sample drawn from students in need of additional reading help as determined by the pretest.

Individualizing a Web-Based Structure Strategy Intervention for Fifth Graders' Comprehension of Nonfiction (Meyer et al., 2011). N schools = 2, N students = 131, Grade = 5, Location = PA.

**Treatment Groups** = All students

question, 2) essential words or key vocabulary related to the unit, 3) knowledge acquisition (appropriate text-based instruction and reading) 4) team-based learning comprehension check, and 5) team-based learning knowledge application. Students in an intervention class received instruction during their regularly scheduled social studies classes. Teachers implemented PACT instruction for 30 full class periods (six to eight weeks). **Randomization** = In participating schools, students were first randomly assigned to classes. Classes were then randomly assigned to treatment or control.

**Treatment Defined** = Treatment students used the *Read Naturally* software for 30–45 min daily, five days per week, for eight weeks. The software offered children their choice of 12 texts. The software offered pronunciation and vocabulary help as needed. Students could move on to the next story only if they independently read the story out loud with no more than three errors. **Randomization** = Students were paired by race, grade, gender, and pretest score and then one student from each pairing was randomly assigned to treatment.

**Treatment Defined** = The ITSS software builds reading comprehension by reading passages of increasing difficulty with students before asking them to read by themselves. The program aims to build familiarity with the following text structures: comparison, problem and

using the constructed latent estimates for each group and their corresponding variances. **Results** = Treatment had a  $0.195\sigma$  (0.077) impact on reading test scores.

**Test Score** = Dynamic Indicators of Basic Early Literacy Skills: Oral Reading Fluency subtest; Expressive Vocabulary Tests; Peabody Picture Vocabulary Test; and the cognitive and achievement batteries of the Woodcock–Johnson III.

**Regression Specification** = Effect sizes were calculated using the average growth between pre and posttest scores. We report the average effect across all outcomes. **Results** = Treatment had a  $0.096\sigma$  (0.221) impact on reading test scores.

**Test Score** = The Gray Silent Reading Test. **Regression Specification** = The effect size was calculated using the average growth between pre and posttest scores. **Results** = Treatment had a  $0.266\sigma$  (0.265) impact on reading test scores.

**Table A4** Curriculum—cont'd

Study	Study Design	Results
<p>used the Intelligent Tutor Structure Strategy (ITSS) computer software. Treatment students received individual lessons from the program. Control students utilized the program's normal curriculum.</p>	<p>solution, cause and effect, sequence, and description. Students in the treatment condition received remediation or enrichment lessons depending on their demonstrated proficiency. Remediation lessons had students read texts of similar or easier complexity, while enrichment lessons had students read the most complex text suited to their ability. The intervention occurred in class over three, 30-min blocks per week.</p> <p><b>Randomization</b> = Students were stratified by pretest scores and elementary school and then randomly assigned to treatment.</p>	
<p>Large-Scale Randomized Controlled Trial with 4th Graders Using Intelligent Tutoring of the Structure Strategy to Improve Nonfiction Reading Comprehension (<a href="#">Wijekumar et al., 2012</a>). N teachers = 130, N students = 2,643, Grades = 4.</p> <p><b>Treatment Groups</b> = Treatment teachers had access to professional development and a web-based intelligent tutoring system (ITSS). Control teachers continued with normal curricula.</p>	<p><b>Treatment Defined</b> = Treatment group received the structure strategy through ITSS. Structure strategy is an approach to improving reading comprehension that focuses on common patterns used by writers to organize texts and organize main ideas. ITSS was designed to deliver instruction within existing English language arts curriculum for one class period a week and provide one on one tutoring. <b>Randomization</b> = Classrooms were randomly assigned to treatment and control within each school. If a school did not have enough classrooms, schools with similar characteristics were grouped and then randomized within that group. All schools volunteered to participate.</p>	<p><b>Test Score</b> = The Gray Silent Reading Test. <b>Regression Specification</b> = Multilevel hierarchical linear model (student, classroom, school) controlling for pretest scores, gender, and school locale. <b>Results</b> = Treatment had a <math>0.10\sigma</math> (<math>0.06</math>) impact on reading test scores.</p>

National Assessment of Title I Interim Report: Volume II: Closing the Reading Gap: First Year Findings from a Randomized Trial of Four Reading Interventions for Striving Readers (Torgesen et al., 2006). N districts = 27, N schools = 50, N students = 772, Grades = 3 and 5.

**Treatment Groups** = Four treatment groups: group one implemented the *Corrective Reading* program, group two implemented the *Failure Free Reading* program, group three implemented the *Spell Read P.A.T.* program, and group four implemented the *Wilson Reading* program. The control group continued with their normal curricula. Sample drawn from students at-risk of developing reading difficulty as determined by pretest scores.

Reading and Language Outcomes of a Multiyear Randomized Evaluation of Transitional Bilingual Education (Slavin et al., 2011). N districts = 6, N schools = 6, N students = 482, Grade = K. **Treatment**

**Groups** = Treatment students participated in a transitional bilingual education program (TBE). Control students utilized a Structured English

**Treatment defined** = All interventions took place for 50 min, five days per week, in groups of three students. The intervention replaced normal reading instruction. The *Spell Read* program builds phonemic awareness via specific sound tasks as well as reading and writing activities. The *Corrective Reading* program uses scripted lessons and rapid exercises to build word identification and fluency. The *Wilson Reading* program uses direct, multisensory, structured reading to build understanding of the structure of language. The *Failure Free* program builds vocabulary through a combination of computer-based lessons, workbook exercises, and teacher-led instruction.

**Randomization** = Schools were sorted into blocks based on the percentage of students eligible for free/reduced lunch. Schools were then stratified by block and, within each block, were randomly assigned to one of the four treatment conditions. Within each school, students were randomly assigned to treatment or control.

**Treatment Defined** = Students in the TBE condition received reading instruction in Spanish for their Kindergarten year. The intervention focused on developing knowledge of the letter sounds, phonics, vocabulary, and concepts of the Spanish language. Most treatment students transitioned to English-language reading classes in 2nd grade. Control utilized the same

**Test Score** = Woodcock Reading Mastery Test-Revised: Word Attack, Word Identification, and Passage Comprehension subtests; Test of Word Reading Efficiency: Phonemic Decoding and Sight Word subtests; The Group Reading and Diagnostic Evaluation.

**Regression Specification** = Two-level hierarchical linear model (student, school) controlling for grade, pretest scores, and block. **Results** = The *Corrective Reading* program had a  $0.148\sigma$  (0.148) impact on reading test scores. The *Failure Free Reading* program had a  $0.048\sigma$  (0.137) impact on reading test scores. The *Spell Read P.A.T.* program had a  $0.179\sigma$  (0.137) impact on reading test scores. The *Wilson Reading* program had a  $0.176\sigma$  (0.147) impact on reading test scores.

**Test Score** = Peabody Picture Vocabulary Test; Woodcock-Johnson: Word Identification, Word Attack, and Passage Comprehension subtests. **Regression**

**Specification** = Effect sizes were calculated using posttest means adjusted for pretest scores. We report the average annual impact across all subtests.

**Results** = Treatment had a  $-0.046\sigma$  (0.070) impact on reading test scores.

**Table A4** Curriculum—cont'd

Study	Study Design	Results
<p>Immersion (SEI) program. Sample composed entirely of students who were Spanish dominant as determined by the pretest.</p>	<p>curriculum as treatment, however all classes were taught in English. All students received regular ESL instruction.</p> <p><b>Randomization</b> = Students were stratified by school and randomly assigned to treatment.</p>	
<p>Segmentation/Spelling Instruction as Part of a First-Grade Reading Program: Effects on Several Measures of Reading (<a href="#">Uhry and Shepherd, 1993</a>). N classrooms = 2, N students = 22, Grade = 1, Location = New York City, NY.</p> <p><b>Treatment Groups</b> = The treatment classroom incorporated phonetic segmentation and spelling techniques into their curriculum. The control classroom retained a reading-based curriculum. Sample drawn from predominately white, middle class, and college educated families.</p>	<p><b>Treatment Defined</b> = All classrooms in the study used the same vocabulary lists. Treatment students were trained to spell phonetically via a series of segmentation, dictation, and computer exercises. The intervention took place over two 20-min sessions per week from October to May in place of normal reading instruction.</p> <p><b>Randomization</b> = Students were stratified by gender and reading ability as determined by the teacher, then randomly assigned to treatment.</p>	<p><b>Test Score</b> = Woodcock Reading Mastery Tests: Word Attack and Word Identification subtests; the Gray Oral Reading Tests. <b>Regression Specification</b> = For each outcome measure, effect sizes were calculated using the average growth between posttest and pretest scores. We report the average effect across all outcome measures.</p> <p><b>Results</b> = Treatment had a <math>1.413\sigma</math> (0.594) impact on reading test scores.</p>
<p>Spatial Temporal Mathematics at Scale: An Innovative and Fully Developed Paradigm to Boost Math Achievement Among All Learners (<a href="#">Rutherford et al., 2010</a>). N schools = 34, Grades = 2–5, Location = CA.</p> <p><b>Treatment Groups</b> = Treatment group implemented the <i>Spatial Temporal Math</i> (<i>ST Math</i>) curriculum. The control group continued their regular curricula. No schools had previous experience with the <i>ST Math</i> curriculum.</p>	<p><b>Treatment defined</b> = The <i>ST Math</i> curriculum utilizes software to present mathematical concepts through a series of pictures and games. The goal of this curriculum is to develop spatial reasoning skills and problem solving techniques. The program was administered for 45 min in class twice weekly.</p> <p><b>Randomization</b> = To be eligible for the study, schools had to be in the lowest three deciles of the Academic Performance Index—a weighted composite of student standardized test scores. Eligible schools who applied to</p>	<p><b>Test Score</b> = California Standards Test: Math subtest. <b>Regression Specification</b> = OLS regression controlling for grade, percent of English Language Learners in each school, percent of students on free/reduced lunch, and the mean test scores of the same grade from the previous year.</p> <p><b>Results</b> = Treatment had a <math>0.290\sigma</math> (0.140) impact on math test scores.</p>

Teaching Children to Become Fluent and Automatic Readers (Kuhn et al., 2006). N schools = 8, N classrooms = 24, N students = 396, Grade = 2. **Treatment**

**Groups** = Two treatment groups: one implemented a repeated-reading curriculum; the other implemented a wide reading curriculum. The control group continued its normal curricula.

Technology's Edge: The Educational Benefits of Computer-Aided Instruction (Barrow et al., 2009). N districts = 3, N schools = 17, N students = 3451. **Treatment**

**Groups** = Treatment classrooms used computer-aided instruction. Control classrooms continued with traditional curricula.

participate were assigned at random to one of two groups: one implemented treatment for second and third graders, the other implemented treatment for fourth and 5th graders.

**Treatment Defined** = In schools implementing a repeated-reading treatment, teachers introduced and discussed a text in class at the start of the week. Students then read the same text approximately four to seven times throughout the week between class time and homework. In schools implementing a wide reading curriculum, teachers introduced and discussed approximately three texts per week, with students re-reading the texts approximately twice between class time and homework.

**Randomization** = Schools were randomly assigned to one of the three groups.

**Treatment Defined** = The study uses a group of computer programs known as I Can Learn. The system is comprised of a software computer package that is designed to deliver instruction through technology on a one-on-one basis to every student. The curricula is designed to meet the National Council of Teachers of Mathematics standards as well as each individual district's course objectives for prealgebra and/or algebra. In addition, the software package also includes a classroom management tool for educators, and the company provides

**Test Score** = Test of Word Reading Efficiency: Significant Word Efficiency subtest; the Gray Oral Reading Test; Wechsler Individual Achievement Test: Reading Comprehension subtest.

**Regression Specification** = Effect sizes were calculated for each outcome measure using posttest means. We report the effect across all outcome measures.

**Results** = The repeated-reading curriculum treatment had a  $0.145\sigma$  ( $0.501$ ) impact on reading test scores. The wide reading curriculum treatment had a  $0.163\sigma$  ( $0.501$ ) impact on reading test scores.

**Test score** = State math tests. **Regression Specification** = OLS regressions controlling for pretest scores, randomization pools, and demographic characteristics. **Results** = Treatment had a  $0.137\sigma$  ( $0.111$ ) impact on math test scores.

**Table A4** Curriculum—cont'd

Study	Study Design	Results
<p>The Effect of Computer Assisted Instruction in Improving Mathematics Performance of Low Achieving Ninth Grade Students (Bailey, 1991). N schools = 1, N teachers = 4, N classes = 4, N students = 46, Grade = 9, Location = Urban high school in Hampton, VA. <b>Treatment Groups</b> = Treatment students were taught the standard curriculum augmented with computer-assisted instruction in their math class. Control students were taught the standard curriculum.</p>	<p>on-site support for administrators and teachers. <b>Randomization</b> = All prealgebra and algebra classes in participating schools were grouped into 60 randomization pools. These pools were defined within a school and typically represent a class period. Within each pool, classes were randomly assigned to treatment and control groups.</p> <p><b>Treatment Defined</b> = Computer-assisted instruction is a method of using computers as a tool to present individualized instructional material. Students in the treatment group used the same textbook as control students. Treatment students also used software that assisted them in learning mathematics skills of concepts, computations, and problem solving. <b>Randomization</b> = The sample of students was identified from students scoring below the 30th percentile on the Iowa Test of Basic Skills mathematics subtest in the previous year and receiving a D or F in their eighth-grade mathematics course. Consent forms were sent to the parents of eligible students. Students who returned the consent form were randomly assigned to treatment or control.</p>	<p><b>Test Score</b> = Test of Achievement and Proficiency: Math subtest. <b>Regression Specification</b> = Effect size was calculated using the average growth between posttest and pretest scores. <b>Results</b> = Treatment had a <math>0.728\sigma</math> (<math>0.304</math>) impact on math test scores.</p>
<p>The Effect of Second-Language Instruction on the Reading Proficiency and General School Achievement of Primary-Grade Children (Potts, 1967). N</p>	<p><b>Treatment Defined</b> = The treatment group received French instruction by the audio-lingual method for 15 min daily over the course of the school year. The control group was given dance</p>	<p><b>Test Score</b> = California Achievement Test. <b>Regression Specification</b> = Effect size was calculated using the posttest means adjusted for language mental age and nonlanguage mental age</p>

classrooms = 4, N students = 80, Grade = 1 and 2, Location = NY. **Treatment Groups** = Treatment students received instruction in French, control students did not.

The Effectiveness of a Program to Accelerate Vocabulary Development in Kindergarten (VOCAB) ([Goodson et al., 2010](#)). N districts = 35, N schools = 65, N teachers = 130, N students = 1,319, Grade = K, Location = Mississippi Delta region. **Treatment Groups** = Treatment students participated in K-PAVE, a kindergarten curriculum designed to enhance students' vocabulary knowledge. Control students continued with their normal curricula.

instruction during this time period. **Randomization** = Students (stratified by gender) and teachers were randomly assigned to the classrooms. Each classroom was then randomly divided in half. Half of the class was assigned to treatment and the other half was assigned to control.

**Treatment Defined** = K-PAVE is built around three components that support the acquisition of vocabulary in young students: (1) instruction on a large set of thematically related target words; (2) interactive book reading to build vocabulary and comprehension skills; and (3) adult-child conversations to build vocabulary and oral language skills. The K-PAVE program was designed as a 24 week supplement to the core language arts program used in each school.

**Randomization** = Participating schools were placed in three blocks based on previous participation in reading initiatives. Within the blocks, schools were matched on school performance classification, percentage of free or reduced-price meal students, percentage of African American students, locale, and region. After being matched, schools were randomly assigned to treatment or control. Furthermore, the researchers randomly selected two consenting kindergarten teachers from treatment schools to collect data from.

as found on the California Test of Mental Maturity. **Results** = Treatment had a  $0.04\sigma$  (0.22) impact on reading test scores.

**Test Score** = The Expressive Vocabulary Test-2. **Regression**

**Specification** = Researchers used a three-level linear hierarchical model (student, teacher, school) controlling for student and school baseline demographics. **Results** = Treatment had a  $0.141\sigma$  (0.052) impact on reading test scores.

**Table A4** Curriculum—cont'd

Study	Study Design	Results
<p>The Effectiveness of Computer Assisted Instruction of Chapter I Students in Secondary Schools (Davidson, 1985). N schools = 1, N students = 67, Grades = 9–12. <b>Treatment Groups</b> = Treatment classes implemented a computer-assisted learning intervention. Control classrooms received no such intervention. Sample drawn from Chapter I students—those who failed to achieve 80 percent of grade-level objectives on local or state standardized tests.</p>	<p><b>Treatment Defined</b> = Treatment students completed mathematics practice problems on a computer for at least 20 min per day for 13 weeks. This intervention used time that would otherwise have been allocated to in-class mathematics exercises.</p> <p><b>Randomization</b> = Students were sorted into five classes by school administrators and then classes were randomly assigned to treatment.</p>	<p><b>Test Score</b> = Metropolitan Achievement Test Battery: Mathematics Instructional subtest. <b>Regression Specification</b> = Effect size was calculated using posttest means adjusted for pretest scores.</p> <p><b>Results</b> = Treatment had a <math>0.121\sigma</math> (<math>1.119</math>) impact on math test scores.</p>
<p>The Effects of Computer Assisted Instruction as a Supplement to Classroom Instruction in Reading Comprehension and Arithmetic (Easterling, 1982). N schools = 3, N students = 72, Grade = 5.</p> <p><b>Treatment Groups</b> = Two treatment conditions: condition one students utilized the <i>SRA Computer Drill and Instruction: Mathematics</i> program; condition two utilized the <i>MicroSystem80</i> reading program. The control group continued with their normal curricula.</p>	<p><b>Treatment Defined</b> = The <i>SRA</i> software offers practice with immediate feedback in basic arithmetic skills. The <i>MicroSystem80</i> software introduces students to the basic rules of logic and drills students on critical reasoning skills. Treatment students worked for a total of 4 h on the computer in addition to normal class time. <b>Randomization</b> = Six boys and six girls were randomly selected from each school. These were matched to another student in the same school on the basis of gender and pretest scores. In two randomly selected schools, pairings were stratified by gender and assigned at random to one of the two treatments. Pairings from the third school served as the control group.</p>	<p><b>Test Score</b> = California Achievement Test. <b>Regression Specification</b> = Effect sizes were calculated using average growth between pre and posttest scores.</p> <p><b>Results</b> = The math instruction treatment had a <math>0.318\sigma</math> (<math>0.301</math>) impact on math test scores and a <math>-0.099\sigma</math> (<math>0.299</math>) impact on reading test scores. The reading instruction treatment had a <math>0.179\sigma</math> (<math>0.299</math>) impact on math test scores and a <math>-0.010\sigma</math> (<math>0.299</math>) impact on reading test scores.</p>

The Enhanced Reading Opportunities (ERO) Study Final Report: The Impact of Supplemental Literacy Courses for Struggling Ninth-grade Readers (Somers et al., 2010). N districts = 10, N schools = 34, N students = 5,595, Grade = 9.

**Treatment Groups** = Treatment one students participated in the Reading Apprenticeship Academic Literacy (RAAL) intervention. Treatment two students participated in the Xtreme Reading intervention. Control students remained in regularly scheduled elective classes.

The Impact of Challenging Geometry and Measurement Units on Achievement of Grade 2 Students (Gavin et al., 2013). N schools = 11, N teachers = 24, N students = 380, Grade = 2, Location = CT, KY, SC, and TX. **Treatment**

**Groups** = Treatment teachers implemented Project M<sup>2</sup> in their classrooms. Control teachers did not.

**Treatment Defined** = The goal of both of the reading interventions is to help struggling adolescent readers develop the strategies and routines used by proficient readers. To do so, each program supports instruction in the following areas: (1) student motivation and engagement; (2) reading fluency; (3) vocabulary; (4) comprehension; (5) phonics and phonemic awareness; and (6) writing.

**Randomization** = Within each district, high schools were randomly assigned to use one of the two supplemental literacy programs. Eligible students within each of the participating high schools were randomly assigned either to enroll in the ERO class or to take one of their school's regularly offered elective classes.

**Treatment Defined** = Project M<sup>2</sup> gave treatment students challenging geometry and measurement units. The purpose is to help primary students learn more complex geometry and measurement concepts in depth. Teachers in treatment group attended a 4-day summer institute and received an additional day of professional development prior to the implementation of each unit. This was a three year intervention.

**Randomization** = Participants were recruited from both lower and higher socioeconomic districts. Teachers were stratified by school and randomly assigned to either treatment or control.

**Test Score** = State test scores. **Regression Specification** = OLS regressions

controlling for students' baseline test scores, whether students were overage at the start of 9th grade, and randomization block fixed-effects. **Results** = The ERO program had a  $0.07\sigma$  (0.035) impact on math test scores and a  $0.11\sigma$  (0.037) impact on reading test scores.

**Test Score** = Iowa Test of Basic Skills: Mathematics Concepts subtest.

**Regression Specification** = Two-level hierarchical linear model (student, class) controlling for pretest scores.

**Results** = Treatment had an impact of  $0.071\sigma$  (0.112) on math test scores.

*Continued*

**Table A4** Curriculum—cont'd

Study	Study Design	Results
<p>The Impact of Collaborative Strategic Reading on the Reading Comprehension of Grade 5 Students in Linguistically Diverse Schools (<a href="#">Hitchcock et al., 2011</a>). N districts = 5, N schools = 26, N students = 1,355, Grade = 5, Location = OK and TX. <b>Treatment Groups</b> = Treatment students participated in Collaborative Strategic Reading (CSR). Control students continued with their normal curricula.</p>	<p><b>Treatment Defined</b> = CSR is a set of instructional strategies designed to improve the reading comprehension of students with diverse abilities. Teachers implement CSR at the classroom level using scaffolded instruction to guide students in the independent use of four comprehension strategies; students apply the strategies to informational text while working in small cooperative learning groups. The goals are to improve reading comprehension and conceptual learning so that academic performance also improves. Treatment teachers received a two day training session on CSR.</p> <p><b>Randomization</b> = Classrooms in participating schools were randomly assigned to CSR treatment or control.</p> <p><b>Treatment Defined</b> = Treatment teachers used a mathematics curriculum that the researchers found successful in a previous correlational study. The curriculum emphasized student practice and teacher presentations through a daily teaching routine the teachers were expected to follow. Treatment teachers attended two 90-min training sessions and received a curriculum manual that they were instructed to read before the start of the experiment. Control teachers received similar training and materials after the completion of the experiment.</p> <p><b>Randomization</b> = Schools were matched by student socioeconomic status and then one school from each pair was randomly assigned to treatment and the other to control.</p>	<p><b>Test Score</b> = The Group Reading Assessment and Diagnostic Evaluation.</p> <p><b>Regression Specification</b> = Two-level hierarchical linear model (student, classroom) controlling for student-level pretest scores, English language learner status, teachers' Spanish fluency, and school fixed-effects. <b>Results</b> = Treatment had an impact of <math>0.05\sigma</math> (0.03) on reading test scores.</p>
<p>The Missouri Mathematics Effectiveness Project: An Experimental Study in Fourth-Grade Classrooms (<a href="#">Good and Grouws, 1979</a>). N schools = 27, N teachers = 40, Grade = 4, Location = Tulsa Public School system. <b>Treatment Groups</b> = Treatment teachers used a new mathematics curriculum. Control teachers continued with their normal curricula.</p>		<p><b>Test Score</b> = Science Research Associates Mathematics test. <b>Regression Specification</b> = Effect size was calculated using average growth between pre and posttest scores. <b>Results</b> = Treatment had a <math>0.648\sigma</math> (0.395) impact on math test scores.</p>

The Relationship Between Supplemental Computer Assisted Mathematics Instruction and Student Achievement ([Manuel, 1987](#)). N schools = 3, N students = 190, Grades = 3–6, Location = Omaha, NE. **Treatment Groups** = Classrooms in treatment one incorporated Computer Curriculum Corporation (CCC) software into their curricula. Classrooms in treatment two incorporated Apple software into their curricula. The control classrooms continued with their normal curricula.

Two-Year Impacts of a Universal School-Based Social-Emotional and Literacy Intervention: An Experiment in Translational Developmental Research ([Jones et al., 2011](#)). N schools = 18, N students = 1,184, Grades = K-3, Location = New York City, NY. **Treatment Groups** = Treatment group implemented the 4R's ("Reading, Writing, Respect, and Resolution") Program. The control group continued the regular curriculum.

**Treatment Defined** = Both treatment programs focused on teaching addition, subtraction, multiplication, division, and problem solving. The lessons were differentiated among students by pretest and Cognitive Skills Index test scores. The CCC software progressively adjusted the difficulty of its assessments based on student performance, and it included limited graphics and no gaming techniques. The Apple software had teachers set the difficulty of assessments, and it incorporated extensive graphics and some gaming techniques. Treatment is in addition to normal class time.

**Randomization** = Students were stratified by grade, gender, and ability (as determined by the Cognitive Skills Index) and then assigned randomly to treatment or control.

**Treatment Defined** = The 4R's program combines academic instruction in language arts with emotional development. The goal of the program is to curb aggression via anger management, listening, assertiveness, cooperation, negotiation, mediation, community building, countering bias, and celebration of differences. **Randomization** = 24 schools agreed to participate and were matched in pairs based on similar demographics. The nine best matched pairs of schools were selected for the study and the other three pairs were reserved as backups. One school from each pairing was assigned at random to treatment.

**Test Score** = California Test of Basic Skills: Mathematics subtest. **Regression Specification** = Average growth in test scores was used to calculate effect sizes.

**Results** = The CCC software treatment had a  $0.066\sigma$  (0.161) impact on math test scores. The Apple software treatment had a  $-0.118\sigma$  (0.230) impact on math test scores.

**Test Score** = New York State standardized assessments of math and reading. **Regression Specification** = Three-level hierarchical linear model (time, student, school) controlling for socioeconomic status, community risk factors, student behavioral risk, gender, race, teacher experience, class size, a survey measure of how burnt out a teacher is, and school fixed-effects. We report annual effects.

**Results** = Treatment had an impact of  $-0.051\sigma$  (0.169) on math test scores and  $-0.012\sigma$  (0.184) on reading test scores.

**Table A4** Curriculum—cont'd

Study	Study Design	Results
<p>Using Enrichment Reading Practices to Increase Reading Fluency, Comprehension, and Attitudes (Reis et al., 2008). N schools = 2, N students = 558, Grades = 3–5.</p> <p><b>Treatment Groups</b> = Treatment teachers incorporated the <i>Schoolwide Enrichment Reading (SEM-R)</i> program into their curricula. Control teachers continued with their normal curricula.</p>	<p><b>Treatment Defined</b> = <i>SEM-R</i> is an enrichment reading program where curriculum is customized based on students' learning styles, needs, and interests. Treatment students participated in 1 h of the school's normal reading program and 1 h of the <i>SEM-R</i> intervention. Control students received 2 h of the school's normal reading program. <b>Randomization</b> = Students and teachers were randomly assigned to treatment or control.</p>	<p><b>Test Score</b> = Iowa Test of Basic Skills: Reading Comprehension subtest.</p> <p><b>Regression Specification</b> = Two-level hierarchical linear model (student, classroom) controlling for oral fluency and school fixed effects. <b>Results</b> = Treatment had an impact of <math>0.28\sigma</math> (<math>0.25</math>) on reading test scores.</p>

## REFERENCES

- Aaronson, D., 1998. Using sibling data to estimate the impact of neighborhoods on children's educational outcomes. *J. Hum. Resour.* 33 (4), 915–946.
- Aaronson, D., Barrow, L., Sander, W., 2007. Teachers and student achievement in the Chicago public high schools. *J. Labor Econ.* 25 (1), 95–135.
- Abdulkadiroglu, A., Angrist, J., Dynarski, S., Kane, T., Pathak, P., 2011. Accountability and flexibility in public schools: evidence from Boston's charters and pilots. *Q. J. Econ.* 126 (2), 699–748.
- Administration for Children and Families, 2006. Preliminary Findings from the Early Head Start Prekindergarten Followup. U.S. Department of Health and Human Services Report, Washington, DC.
- Ainsworth, J., 2002. Why does it take a village? The mediation of neighborhood effects on educational achievement. *Soc. Forces* 81 (1), 117–152.
- Akerlof, G., 1978. The economics of "Tagging" as applied to the optimal income tax, welfare programs, and manpower planning. *Am. Econ. Rev.* 68 (1), 8–19.
- Alexander, K., Entwistle, D., Olson, L., 2001. Schools, achievement, and inequality: a seasonal perspective. *Educ. Eval. Policy Anal.* 23 (2), 171–191.
- Allington, R., McGill-Franzen, A., Camilli, G., Williams, L., Graf, J., Zeig, J., Zmach, C., Nowak, R., 2010. Addressing summer reading setback among economically disadvantaged elementary students. *Read. Psychol.* 31 (5), 411–427.
- Allor, J., McCathren, R., 2004. The efficacy of an early literacy tutoring program by college students. *Learn. Disabil. Res. Pract.* 19 (2), 116–129.
- Almond, D., Currie, J., 2010. Human capital development before age five. *Handb. Labor Econ.* 4b, 1315–1486.
- Al Otaiba, S., Connor, C., Folsom, J., Greulich, L., Meadows, J., Li, Z., 2011. Assessment data-informed guidance to individualize kindergarten reading instruction: findings from a cluster-randomized control field trial. *Elem. Sch. J.* 111 (4), 535–560.
- Alpert, J., 1975. Teacher behavior and pupil performance: reconsideration of the mediation of Pygmalion effects. *J. Educ. Res.* 69 (2), 53–57.
- Angrist, J., Bettinger, E., Bloom, E., King, E., Kremer, M., 2002. Vouchers for private schooling in Colombia: evidence from a randomized natural experiment. *Am. Econ. Rev.* 92 (5), 1535–1558.
- Angrist, J., Bettinger, E., Kremer, M., 2006. Long-term educational consequences of secondary school vouchers: evidence from administrative records in Colombia. *Am. Econ. Assoc.* 96 (3), 847–862.
- Angrist, J., Dynarski, S., Kane, T., Pathak, P., Walters, C., 2011. Who Benefits from KIPP? IZA Discussion Paper no. 5690.
- Angrist, J., Lang, D., Oreopoulos, P., 2009. Incentives and services for college achievement: evidence from a randomized trial. *Am. Econ. J. Appl. Econ.* 1 (1), 136–163.
- Angrist, J., Lavy, V., 2009. The effects of high stakes high school achievement awards: evidence from a randomized trial. *Am. Econ. Rev.* 99 (4), 1384–1414.
- Angrist, J., Pathak, P., Walters, C., 2013. Explaining charter school effectiveness. *Am. Econ. J. Appl. Econ.* 5 (4), 1–27.
- Aronson, J., Fried, C., Good, C., 2002. Reducing the effects of stereotype threat on African American college students by shaping theories of intelligence. *J. Exp. Soc. Psychol.* 38, 113–125.
- Arvans, R., 2009. Improving Reading Fluency and Comprehension in Elementary Students Using Read Naturally. Dissertation submitted to Western Michigan University.
- Assel, M., Landry, S., Swank, P., Gunnewig, S., 2007. An evaluation of curriculum, setting, and mentoring on the performance of children enrolled in pre-kindergarten. *Read. Writ.* 20 (5), 463–494.
- Attewell, P., Battle, J., 1999. Home computers and school performance. *Inf. Soc.* 15, 1–10.
- Aud, S., Hussar, W., Kena, G., Bianco, K., Frohlich, L., Kemp, J., Tahan, K., 2011. The Condition of Education 2011. U.S. Department of Education, National Center for Education Statistics, Washington, DC.
- Avvisati, F., Gurgand, M., Guyon, N., Maurin, E., 2014. Getting parents involved: a field Experiment in deprived schools. *Rev. Econ. Stud.* 81 (1), 57–83.

- Bailey, T., 1991. The Effect of Computer-Assisted Instruction in Improving Mathematics Performance of Low-Achieving Ninth-Grade Students. Dissertation submitted to The College of William and Mary.
- Baker, G., 2002. Distortion and risk in optimal incentive contracts. *J. Hum. Resour.* 37 (4), 728–751.
- Baker, A., Piotrkowski, C., Brooks-Gunn, J., 1998. The effects of the Home Instruction Program for Preschool Youngsters (HIPPY) on children's school performance at the end of the program and one year later. *Early Child. Res. Q.* 13 (4), 571–588.
- Baker, S., Gersten, R., Keating, T., 2000. When less may be more: a 2-year longitudinal evaluation of a volunteer tutoring program requiring minimal training. *Read. Res. Q.* 35 (4), 494–519.
- Barlevy, G., Neal, D., 2012. Pay for percentile. *Am. Econ. Rev.* 102 (5), 1805–1831.
- Barnett, S., Jung, K., Yarosz, D., Thomas, J., Hornbeck, A., Stechuk, R., Burns, S., 2008. Educational effects of the tools of the mind curriculum: a randomized trial. *Early Child. Res. Q.* 23 (3), 299–313.
- Barrera-Osorio, F., Bertrand, M., Linden, L., Perez-Calle, F., 2011. Improving the design of conditional transfer programs: evidence from a randomized education experiment in Colombia. *Am. Econ. J. Appl. Econ.* 3 (2), 167–195.
- Barrow, L., Markman, L., Rouse, C., 2009. Technology's edge: the educational benefits of computer-aided instruction. *Am. Econ. J. 1* (1), 52–74.
- Becker, G., 1995. Human Capital and Poverty Alleviation. Human Resource and Operations Policy, World Bank. Working Paper no. 52.
- Behrman, J., Sengupta, P., Todd, P., 2001. Progressing through PROGRESA: An Impact Assessment of a School Subsidy Experiment. International Food Policy Research Institute, Washington, DC.
- Behrman, J., Sengupta, P., Todd, P., 2005. Progressing through PROGRESA: an impact assessment of a school subsidy experiment in rural Mexico. *Econ. Dev. Cult. Change* 54 (1), 237–275.
- Berger, A., Turk-Bicakci, L., Garet, M., Song, M., Knudson, J., Haxton, C., Zeiser, K., Hoshen, G., Ford, J., Stephan, J., 2013. Early College, Early Success: Early College High School Initiative Impact Study. American Institutes for Research, Washington, DC.
- Bettinger, E., 2012. Paying to learn: the effect of financial incentives on elementary school test scores. *Rev. Econ. Statistics* 94 (3), 686–698.
- Bierman, K., Coie, J., Dodge, K., Greenberg, M., Lochman, J., McMahon, R., Pinderhughes, E., 2002. Evaluation of the first 3 years of the Fast Track prevention trial with children at high risk for adolescent conduct problems. *J. Abnorm. Child Psychol.* 30 (1), 19–35.
- Bierman, K., Domitrovich, C., Nix, R., Gest, S., Welsh, J., Greenberg, M., Blair, C., Nelson, K., Gill, S., 2008. Promoting academic and social-emotional school readiness: the head start REDI program. *Child Dev.* 79 (6), 1802–1817.
- Bifulco, R., Cobb, C., Bell, C., 2009. Can interdistrict choice boost student achievement? The case of Connecticut's interdistrict magnet school program. *Educ. Eval. Policy Anal.* 31 (4), 323–345.
- Bill and Melinda Gates Foundation, 2014. Teacher's Know Best: Teachers' Views on Professional Development. Bill and Melinda Gates Foundation, Seattle, WA.
- Blachman, B., 1987. An alternative classroom Reading program for learning disabled and other low-achieving children. In: Bowler, R. (Ed.), *Intimacy with Language: A Forgotten Basic in Teacher Education*. Orton Dyslexia Society, Baltimore, MD.
- Blachman, B., Tangel, D., Wynne Ball, E., Black, R., McGraw, C., 1999. Developing phonological awareness and word recognition skills: a two-year intervention with low-income, inner-city children. *Read. Writ.* 11 (3), 239–273.
- Blachman, B., Schatschneider, C., Fletcher, J., Francis, D., Clonan, S., Shaywitz, B., Shaywitz, S., 2004. Effects of intensive reading remediation for second and third graders and a 1-year follow-up. *J. Educ. Psychol.* 96 (3), 444–461.
- Black, A., Somers, M., Doolittle, F., Unterman, R., Grossman, J., Warner, E., 2009. The Evaluation of Enhanced Academic Instruction in After-School Programs: Final Report. U.S. Department of Education, National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, Washington, DC.
- Bloom, N., Lemos, R., Sadun, R., Van Reenen, J., 2015. Does management matter in schools? *Econ. J.* 125 (584), 647–674.

- Boller, K., Vogel, C., Johnson, A., Novak, T., James-Burdumy, S., Crozier, L., et al., 2004. Using Television as a Teaching Tool: The Impacts of Ready to Learn Workshops on Parents, Educators, and the Children in Their Care (No. PR04-63). Mathematica Policy Research, Inc.
- Borman, G., Slavin, R., Cheung, A., Chamberlain, A., Madden, N., Chambers, B., 2007. Final Reading outcomes of the national randomized field trial of success for all. *Am. Educ. Res. J.* 44 (3), 701–731.
- Borman, G., Dowling, N., Schneck, C., 2008. A multisite cluster randomized field trial of open court reading. *Educ. Eval. Policy Anal.* 30 (4), 389–407.
- Borman, G., Benson, J., Overman, L., 2009. A randomized field trial of the Fast ForWord Language computer-based training program. *Educ. Eval. Policy Anal.* 31 (1), 82–106.
- Bos, J., Sanchez, R., Tseng, F., Rayyes, N., Ortiz, L., Sinicrope, C., 2012. Evaluation of Quality Teaching for English Learners (QTEL) Professional Development. U.S. Department of Education, National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, Washington, DC.
- Bowers, W., 1972. An Evaluation of a Pilot Program in Reading for Culturally Disadvantaged First Grade Students. Dissertation submitted to the University of Tulsa.
- Boyd, D., Goldhaber, D., Lanjford, H., Wyckoff, J., 2007. The effect of certification and preparation on teacher quality. *Future Child.* 17 (1), 45–68.
- Brandt, C., Meyers, C., Molefe, A., 2013. The Impact of Emints Professional Development on Teacher Instruction and Student Achievement: Year 1 Report. American Institutes for Research.
- Broh, B., 2004. Racial/Ethnic Achievement Inequality: Separating School and Non-School Effects Through Seasonal Comparisons. Dissertation submitted to Ohio State University, Athens, OH.
- Brooks-Gunn, J., Duncan, G., 1997. The effects of poverty on children. *Future Child.* 7 (2), 55–71.
- Brooks-Gunn, J., Klebanov, P., Smith, J., Duncan, G., Lee, K., 2003. The black-white test score gap in young children: contributions of test and family characteristics. *Appl. Dev. Sci.* 7 (4), 239–252.
- Brooks-Gunn, J., Liaw, F.-R., Klebanov, P.K., 1992. Effects of early intervention on cognitive function of low birth weight preterm infants. *J. Pediatr.* 120 (3), 350–359.
- Brooks-Gunn, J., McCarton, C., Casey, P., McCormick, M., Bauer, C., Bernbaum, J., Tyson, J., Swanson, M., Bennett, F., Scott, D., et al., 1994. Early intervention in low-birth-weight premature infants. Results through age 5 years from the Infant Health and Development Program. *JAMA* 272 (16), 1257–1262.
- Campbell, C., Brigman, G., 2005. Closing the achievement gap: a structured approach to group counseling. *J. Spec. Group Work* 30 (1), 67–82.
- Campbell, P., Malkus, N., 2011. The impact of elementary mathematics coaches on student achievement. *Elem. Sch. J.* 111 (3), 430–454.
- Campbell, F., Ramey, C., 1994. Effects of early intervention on intellectual and academic achievement: a follow-up study of children from low-income families. *Child Dev.* 65 (2), 684–698.
- Campuzano, L., Dynarski, M., Agodini, R., Rall, K., 2009. Effectiveness of Reading and Mathematics Software Products: Findings from Two Students Cohorts. U.S. Department of Education, National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, Washington, DC.
- Cantrell, S., Fullerton, J., Kane, T., Staiger, D., 2008. National Board Certification and Teacher Effectiveness: Evidence from a Random Assignment Experiment. NBER Working Paper no. 14608.
- Carlson, D., Geoffrey, B., Michelle, R., 2011. A multistate district-level cluster randomized trial of the impact of data-driven reform on reading and mathematics achievement. *Educ. Eval. Policy Anal.* 33 (3), 378–398.
- Carneiro, P., Heckman, J., 2003. Human Capital Policy. NBER Working Paper no. 9495.
- Carpenter, T., Fennema, E., Peterson, P., Chiang, C., Loef, M., 1989. Using knowledge of children's mathematics thinking in classroom teaching: an experimental study. *Am. Educ. Res. J.* 26 (4), 499–531.
- Center for Research on Education Outcomes (CREDO), 2013. National Charter School Study. Center for Research on Education Outcomes, Stanford, CA.
- Center, Y., Wheldall, K., Freeman, L., Outhred, L., McNaught, M., 1995. An evaluation of reading recovery. *Read. Res. Q.* 30 (2), 240–263.

- Charity, A., Scarborough, H., Griffin, D., 2004. Familiarity with school english in African American children and its relation to early reading achievement. *Child Dev.* 75 (5), 1340–1356.
- Chase-Lansdale, L., Gordon, R., 1996. Economic hardship and the development of five-and six-year-olds: neighborhood and regional perspectives. *Child Dev.* 67 (6), 3338–3367.
- Chase-Lansdale, L., Gordon, R., Brooks-Gunn, J., Klehanov, P.K., 1997. Neighborhood and family influences on the intellectual and behavioral competence of preschool and early school-age children. In: Brooks-Gunn, J., Duncan, G., Lawrence Aber, J. (Eds.), *Neighborhood Poverty: Context and Consequences for Children*, vol. 1. Russel Sage, New York, pp. 79–118.
- Chenoweth, K., 2007. "It's Being Done": Academic Success in Unexpected School. Harvard Education Press, Cambridge, MA.
- Chetty, R., Friedman, J., Hilger, N., Saez, E., Whitmore Schanzenbach, D., Yagan, D., 2011. How does your kindergarten classroom affect your earnings? Evidence from project STAR. *Q. J. Econ.* 126 (4), 1593–1660.
- Chetty, R., Friedman, J., Rockoff, J., 2014. Measuring the impacts of teachers II: teacher value-added and student outcomes in adulthood. *Am. Econ. Rev.* 104 (9), 2633–2679.
- Chetty, R., Hendren, N., Katz, L., 2016. The effects of exposure to better neighborhoods on children: new evidence from the Moving to Opportunities experiment. *Am. Econ. Rev.* 106 (4), 855–902.
- Christ, T., Davie, J., 2009. Empirical evaluation of Read Naturally effects: a randomized control trial (RCT) (Unpublished journal article).
- Clark, C., Walberg, H., 1968. The influence of massive rewards on reading achievements in potential urban school dropouts. *Am. Educ. Res. J.* 5 (3), 305–310.
- Clark, M., Chiang, H., Silva, T., McConnell, S., Sonnenfeld, K., Erbe, A., 2013. The Effectiveness of Secondary Math Teachers from Teach for America and the Teaching Fellows Programs. National Center for Educational Evaluation and Regional Assistance, Washington, DC.
- Cohen, J., 1965. Some statistical issues in psychological research. In: Wolman, B. (Ed.), *Handbook of Clinical Psychology*. McGraw-Hill, New York.
- Cohen, D., Hill, H., 2001. Learning Policy: When State Education Reform Works. Yale University Press, New Haven, CT.
- Cohen, G., Garcia, J., Apfel, N., Master, A., 2006. Reducing the racial achievement gap: a social-psychological intervention. *Science* 313 (5791), 1307–1310.
- Cohen, G., Garcia, J., Purdie-Vaughns, V., Apfel, N., Brzustoski, P., 2009. Recursive processes in self-affirmation: intervening to close the minority achievement gap. *Science* 324 (5925), 400–403.
- Coie, J., Krebiel, G., 1984. Effects of academic tutoring on the social status of low-achieving, socially rejected children. *Child Dev.* 55 (4), 1465–1478.
- Cole, D., 1992. The Effects of a One-Year Staff Development Program on the Achievement Test Scores of Fourth-Grade Students. Dissertation submitted to the University of Mississippi.
- Coleman, J., Campbell, E., Hobson, C., McPartland, J., Mood, A., Weinfeld, F., York, R., 1966. Equality of Educational Opportunity. U.S. Department of Health, Education, and Welfare, Washington, DC.
- Compton, G., 1992. The Reading Connection: A Leadership Initiative Designed to Change the Delivery of Educational Services to At-Risk Children. Dissertation submitted to Western Michigan University.
- Constantine, J., Player, D., Silva, T., Hallgren, K., Grider, M., Deke, J., 2009. An Evaluation of Teachers Trained Through Different Routes to Certification. National Center for Education Evaluation and Regional Assistance, Washington, DC.
- Cook County Department of Public Aid, 1969. Project Breakthrough: A Responsive Environment Field Experiment with Pre-School Children from Public Assistance Families. U.S. Department of Health, Education Welfare.
- Cook, T., Habib, F., Phillips, M., Settersen, R., Shagle, S., Degirmencioglu, S., 1999. Comer's School Development Program in Prince George's County, Maryland: a theory-based evaluation. *Am. Educ. Res. J.* 36 (3), 543–597.
- Cook, P., Dodge, K., Farkas, G., Fryer, R., Guryan, J., Ludwig, J., Mayer, S., Pollack, H., Steinberg, L., 2014. The (Surprising) Efficacy of Academic and Behavioral Intervention with Disadvantaged Youth: Results from a Randomized Experiment in Chicago. NBER Working Paper no. 19862.

- Cooper, H., Nye, B., Lindsay, J., Greathouse, S., 1996. The effects of summer vacation on achievement test scores: a narrative and meta-analytic review. *Rev. Educ. Res.* 66 (3), 227–268.
- Corcoran, S., Evans, W., Schwab, R., 2004. Changing labor-market opportunities for women and the quality of teachers, 1957–2000. *Am. Econ. Rev.* 94 (2), 729–760.
- Cosgrove, M., Fountain, C., Wehry, S., Wood, J., Kasten, K., 2006. Randomized field trial of an early literacy curriculum and instructional support system. In: Paper Presented at the Annual Meeting of the Educational Research Association. San Francisco, California.
- Courtney, M., Zinn, A., Zielewski, E., Bess, R., Malm, K., 2008. Evaluation of the Early Start to Emancipation Preparation—tutoring Program Los Angeles County, California: Final Report. The Urban Institute, Washington, DC.
- Cowen, J., 2008. School choice as a latent variable: estimating “complier average causal effect” of vouchers in Charlotte. *Policy Stud. J.* 36 (2), 301–315.
- Cullen, J.B., Jacob, B., Levitt, S., 2006. The effect of school choice on participants: evidence from randomized lotteries. *Econometrica* 74 (5), 1191–1230.
- Cullen, J.B., Levitt, S., Robertson, E., Sadoff, S., 2013. What can be done to improve struggling high schools? *J. Econ. Perspect.* 27 (2), 133–152.
- Cunha, F., Heckman, J., 2007. The technology of skill formation. *Am. Econ. Rev.* 97 (2), 31–47.
- Cunha, F., Heckman, J., 2010. Investing in Our Young People. NBER Working Paper no. 16201.
- Currie, J., Thomas, D., 1995. Does head start make a difference. *Am. Econ. Rev.* 85 (3), 341–364.
- Curto, V., Fryer, R., 2014. The potential of urban boarding schools for the poor. *J. Labor Econ.* 32 (1), 65–93.
- Davidson, R., 1985. The Effectiveness of Computer Assisted Instruction of Chapter 1 Students in Secondary Schools. Dissertation submitted to the University of Tennessee, Knoxville.
- Davis-Kean, P., 2005. The influence of parent education and family income on child achievement: the indirect role of parental expectations and the home environment. *J. Fam. Psychol.* 19 (2), 294–304.
- de la Rica, S., 2011. Social and labor market integration of ethnic minorities in Spain. In: Kahanec, M., Zimmerman, K. (Eds.), *Ethnic Diversity in European Labor Markets: Challenges and Solutions*. Edward Elgar Publishing, Cheltenham, UK, pp. 268–282.
- Deaton, A., 2010. Instruments, randomization, and learning about development. *J. Econ. Literature* 48 (2), 424–455.
- Dee, T., Jacob, B., 2011. The impact of No child left behind on student achievement. *J. Policy Anal. Manag.* 30 (3), 418–446.
- Deming, D., Cohodes, S., Jennings, J., Jencks, C., 2013. School Accountability, Postsecondary Attainment and Earnings. NBER Working Paper no. 19444.
- DerSimonian, R., Laird, N., 1986. Meta-analysis in clinical trials. *Control Clin. Trials* 7 (3), 177–188.
- Dobbie, W., Fryer, R., 2011. Are high-quality schools enough to increase achievement among the poor? Evidence from the Harlem children's zone. *Am. Econ. J. Appl. Econ.* 3 (3), 158–187.
- Donaldson, M., Johnson, S.M., 2010. The price of misassignment: the role of teaching assignments in teach for America teacher's exit from low-income schools and the teaching profession. *Educ. Eval. Policy Anal.* 32 (2), 299–323.
- Drummond, K., Chinen, M., Duncan, T., Miller, H., Fryer, L., Zmach, C., Culp, K., 2011. Impact of the Thinking Reading Software Program on Grade 6 Reading, Vocabulary, Comprehension, Strategies, and Motivation. U.S. Department of Education, National Center for Education Evaluation and Regional Assistance.
- Duckworth, A., Kirby, T., Gollwitzer, A., Oettingen, G., 2013. From fantasy to action: mental contrasting with implementation intentions (MCII) improves academic performance in children. *Soc. Psychol. Personality Sci.* 4, 745–753.
- Duflo, E., Hanna, R., Ryan, S., 2012. Incentives work: getting teachers to come to school. *Am. Econ. Rev.* 102 (4), 1241–1278.
- Duncan, G., Brooks-Gunn, J., Klebanov, P.K., 1994. Economic deprivation and early childhood development. *Child Dev.* 65 (2), 296–318.
- Duncan, G., Magnuson, K., 2005. Can family socioeconomic resources account for racial and ethnic test score gaps? *Future Child.* 15 (1), 35–54.

- Dunstan, W., 2010. Ancient Rome. Rowman and Littlefield, Lanham, MD.
- Easterling, B., 1982. The Effects of Computer Assisted Instruction as a Supplement to Classroom Instruction in Reading Comprehension and Arithmetic. Dissertation submitted to North Texas State University.
- Evans, G., 2004. The environment of childhood poverty. *Am. Psychol.* 59 (2), 77–92.
- Fairlie, R., 2005. The effects of home computers on school enrollment. *Econ. Educ. Rev.* 24 (5), 533–547.
- Fairlie, R., Beltran, D., Das, K., 2010. Home computers and educational outcomes: evidence from the NLSY97 and CPS. *Econ. Inq.* 48 (3), 771–792.
- Fairlie, R., Robinson, J., 2013. Experimental evidence on the effects of home computers on academic achievement among school children. *Am. Econ. J. Appl. Econ.* 5 (3), 211–240.
- Fantuzzo, J., Davis, G., Ginsburg, M., 1995. Effects of parent involvement in isolation or in combination with peer tutoring on student self-concept and mathematics achievement. *J. Educ. Psychol.* 87 (2), 272–281.
- Farver, J., Lonigan, C., Eppe, S., 2009. Effective early literacy skill development for young Spanish-speaking English language learners: an experimental study of two methods. *Child Dev.* 80 (3), 703–719.
- Feng, L., 2010. Hire today, gone tomorrow: new teacher classroom assignments and teacher mobility. *Educ. Finance Policy* 5 (3), 278–316.
- Fiorini, M., 2010. The effect of home computer use on children's cognitive and non-cognitive skills. *Econ. Educ. Rev.* 29 (1), 55–72.
- Fischel, J., Bracken, S., Fuchs-Eisenberg, A., Spira, E., Katz, S., Shaller, G., 2007. Evaluation of curricular approaches to enhance preschool early literacy skills. *J. Lit. Res.* 39 (4), 471–501.
- Fleischman, H., Hopstock, P., Pelczar, M., Shelley, B., 2010. Highlights from PISA 2009: Performance of U.S. 15-year-old Students in Reading, Mathematics, and Science Literacy in an International Context. U.S. Department of Education, Washington, DC.
- Fletcher, J., Lyon, R., 1998. Reading: a research-based approach. In: What's Gone Wrong in America's Classrooms. Hoover Institution Press, Stanford, CA.
- Fletcher, J., Hawley, D., Piele, P., 1990. Costs, effects, and utility of microcomputer assisted instruction in the classroom. *Am. Educ. Res. J.* 27 (4), 783–806.
- Foorman, B., Moats, L., 2004. Conditions for sustaining research-based practices in early reading instruction. *Remedial Spec. Educ.* 25 (1), 51–60.
- Friedman, M., 1955. The role of government in public education. In: Solo, R. (Ed.), Economics and the Public Interest. University of Rutgers Press, New Brunswick, NJ.
- Fryer, R., 2010. Racial inequality in the 21st century: the declining significance of discrimination. *Handb. Labor Econ.* 4 (B), 855–971.
- Fryer, R., 2011. Financial incentives and student achievement: evidence from trials. *Q. J. Econ.* 126 (4), 1755–1798.
- Fryer, R., 2013a. Information and Student Achievement: Evidence from a Cellular Phone Experiment. NBER Working Paper no. 19113.
- Fryer, R., 2013b. Teacher incentives and student achievement: evidence from New York city public schools. *J. Labor Econ.* 31 (2), 373–427.
- Fryer, R., 2014a. Injecting charter school best practices into traditional public schools: evidence from field experiments. *Q. J. Econ.* 129 (3), 1355–1407.
- Fryer, R., 2014b. Teacher incentives and student achievement: evidence from New York city public schools. *Q. J. Econ.* 129 (3), 1355–1407.
- Fryer, R., Dobbie, W., 2013. Getting beneath the veil of effective schools: evidence from New York city. *Am. Econ. J. Appl. Econ.* 5 (4), 28–60.
- Fryer, R., Holden, R., 2013. Multitasking, Dynamic Complementaries, and Incentives: A Cautionary Tale (Working Paper).
- Fryer, R., Levitt, S., 2004. Understanding the black-white test score gap in the first two years of school. *Rev. Econ. Statistics* 86 (2), 447–464.
- Fryer, R., Levitt, S., 2006. The black-white test score gap through third grade. *Am. Law Econ. Rev.* 8 (2), 249–281.
- Fryer, R., Levitt, S., 2013. Testing for racial differences in the mental ability of young children. *Am. Econ. Rev.* 103 (2), 981–1005.

- Fryer, R., Levitt, S., List, J., 2015a. Parental Incentives and Early Childhood Achievement: A Field Experiment in Chicago Heights (Unpublished working paper).
- Fryer, R., Levitt, S., List, J., Sadoff, S., 2015b. Enhancing the Efficacy of Teacher Incentives through Loss Aversion: A Field Experiment. NBER Working Paper no. 18237.
- Fuchs, T., Woessmann, L., 2004. Computers and Student Learning: Bivariate and Multivariate Evidence on the Availability and Use of Computers at Home and at School. CESifo Working Paper no. 1321.
- Fuchs, L., Fuchs, D., Kazdan, S., Allen, S., 1999. Effects of Peer-Assisted Learning Strategies in reading with and without training in elaborated help giving. *Elem. Sch. J.* 99 (3), 201–219.
- Fuchs, L., Fuchs, D., Karns, K., 2001. Enhancing kindergarteners' mathematical development: effects of peer-assisted learning strategies. *Elem. Sch. J.* 101 (5), 495–510.
- Fuchs, L., Fuchs, D., Yazdian, L., Powell, S., 2002. Enhancing first-grade children's mathematical development with Peer-Assisted Learning Strategies. *Sch. Psychol. Rev.* 31 (4), 569–583.
- Fuchs, L., Compton, D., Fuchs, D., Paulsen, K., Bryant, J., Hamlett, C., 2005. The prevention, identification, and cognitive determinants of math difficulty. *J. Educ. Psychol.* 97 (3), 493–513.
- Garber, H., 1988. The Milwaukee Project: Preventing Mental Retardation in Children at Risk. National Institute of Handicapped Research, Washington, DC.
- Garces, E., Thomas, D., Currie, J., 2002. Longer-term effects of head start. *Am. Econ. Rev.* 92 (4), 999–1012.
- Garet, M., Cronen, S., Eaton, M., Kurki, A., Jones, W., Uekawa, K., Falk, A., 2008. The Impact of Two Professional Development Interventions on Early Reading Instruction and Achievement. National Center for Education Evaluation and Regional Assistance, Washington, DC.
- Garet, M., Porter, A., Desimone, L., Birman, B., Yoon, K.S., 2001. What makes professional development effective? Results from a national sample of teachers. *Am. Educ. Res. J.* 38 (4), 915–945.
- Gavin, M., Casa, T., Andelson, J., Firmender, J., 2013. The impact of challenging geometry and measurement units on the achievement of grade 2 students. *J. Res. Math. Educ.* 44 (3), 478–509.
- Gersten, R., Dimino, J., Jayanthi, M., Kim, J., Santoro, L., 2010. Teacher Study Group: impact of the professional development model on reading instruction and student outcomes in first grade classrooms. *Am. Educ. Res. J.* 47 (3), 694–739.
- Glass, G., Smith, M.L., 1978. Meta-analysis of Research on the Relationship of Class-size and Achievement. Far West Laboratory for Educational Research and Development, San Francisco, CA.
- Glassman, P., 1989. A Study of Cooperative Learning in Mathematics, Writing and Reading as Implemented in the Intermediate Grades: A Focus upon Achievement, Attitudes, and Self-Esteem by Gender, Race, and Ability Group. Dissertation submitted to Hofstra University.
- Glazerman, S., Mayer, D., Decker, P., 2006. Alternative routes to teaching: the impacts of teach for America on student achievement and other outcomes. *J. Policy Anal. Manag.* 25 (1), 75–96.
- Glazerman, S., McKie, A., Carey, N., 2009. An Evaluation of the Teacher Advancement Program (TAP) in Chicago: Year One Impact Report. Mathematica Policy Research, Princeton, NJ.
- Glazerman, S., Protik, A., The, B., Bruch, J., Max, J., 2013. Transfer Incentives for High-Performing Teachers: Final Results from a Multisite Randomized Experiment. National Center for Education Evaluation and Regional Assistance, Washington, DC.
- Gleason, P., Clark, M., Tuttle, C., Dwoyer, E., 2010. The Evaluation of Charter School Impacts. National Center for Education Evaluation and Regional Assistance, Washington, DC.
- Glewwe, P., Ilias, N., Kremer, M., 2010. Teacher incentives. *Am. Econ. J. Appl. Econ.* 2 (3), 205–227.
- Good, T., Grouws, D., 1979. The Missouri Mathematics Effectiveness Project: an experimental study in fourth-grade classrooms. *J. Educ. Psychol.* 71 (3), 355–362.
- Goodson, B., Wolf, A., Bell, S., Turner, H., Finney, P., 2010. The Effectiveness of a Program to Accelerate Vocabulary Development in Kindergarten (VOCAB). U.S. Department of Education, National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, Washington, DC.
- Gray, S., Klaus, R., 1970. The early training project: a seventh-year report. *Child Dev.* 41, 909–924.
- Greene, J., Peterson, P., Du, J., 1999. Effectiveness of school choice: the Milwaukee experiment. *Educ. Urban Soc.* 31 (2), 190–213.

- Greenwood, C., Delquadri, J., Hall, R., 1989. Longitudinal effects of classwide peer tutoring. *J. Educ. Psychol.* 81, 371–383.
- Gunn, B., Smolkowski, K., Biglan, A., Black, C., Blair, J., 2005. Fostering the development of reading skill through supplemental instruction: results for Hispanic and non-Hispanic students. *J. Spec. Educ.* 39 (2), 66–85.
- Guryan, J., Kim, J., Quinn, D., 2014. Does Reading During the Summer Build Reading Skills? Evidence from a Randomized Experiment in 463 Classrooms. NBER Working Paper No. 20689.
- Hahn, A., Leavitt, T., Aaron, P., 1994. Evaluation of the Quantum Opportunities Program (QOP) Did the Program Work? Brandeis University Center for Human Resources, Waltham, MA.
- Halpern-Felsher, B., Connell, J.P., Beale Spencer, M., Lawrence Aber, J., Duncan, G.P., Elizabeth, C., Crichlow, W.E., Usinger, P.A., Cole, S.P., Allen, LaR., Seidman, E., 1997. Neighborhood and family factors predicting educational risk and attainment in African American and white children and adolescents. In: Brooks-Gunn, J., Duncan, G., Aber, L. (Eds.), *Neighborhood Poverty, Volume I: Context and Consequences for Children*. Russel Sage Foundation, New York.
- Hamilton, G., Freedman, S., Gennetian, L., Michalopoulos, C., Walter, J., Adams-Ciardullo, D., Gassman-Pines, A., 2001. National Evaluation of Welfare-to-Work Strategies. U.S. Department of Health and Human Services, Washington, DC.
- Hanushek, E., 1979. Conceptual and empirical issues in the estimation of educational production functions. *J. Hum. Resour.* 14 (3), 351–388.
- Harrington, M., 1982. *The Other America: Poverty in the United States*. Touchstone, New York, NY.
- Harrison, G., List, J., 2004. Field experiments. *J. Econ. Literature* 42 (4), 1009–1055.
- Hatton, T., 2011. The social and labor market outcomes of ethnic minorities in the UK. In: Kahanec, M., Zimmerman, K. (Eds.), *Ethnic Diversity in European Labor Markets: Challenges and Solutions*. Edward Elgar Publishing, Cheltenham, UK, pp. 283–306.
- Heckman, J., 2008. Role of income and family influence on child outcomes. *Ann. N.Y. Acad. Sci.* 1136, 307–323.
- Heckman, J., Hyeok Moon, S., Pinto, R., Savelyev, P., Yavitz, A., 2009. A Reanalysis of the High/Scope Perry Preschool Program. University of Chicago, Department of Economics (Unpublished Manuscript).
- Heckman, J., Hyeok Moon, S., Pinto, R., Savelyev, P., Yavitz, A., 2010. The rate of return to the high/ scope perry preschool program. *J. Public Econ.* 94 (1–2), 114–128.
- Heckman, J., Kautz, T., 2013. Fostering and Measuring Skills: Interventions that Improve Character and Cognition. NBER Working Paper no. 19656.
- Hedges, L., 1981. Distribution theory for glass's estimator of effect sizes and related estimators. *J. Educ. Behav. Statistics* 6 (2), 107–128.
- Heuys, B., 1978. *Summer Learning and the Effects of Schooling*. Academic Press, Orlando, FL.
- Heuys, B., 1987. Schooling and cognitive development. *Child Dev.* 58 (5), 1151–1160.
- Hill, H., 2007. Learning in the teaching workforce. *Future Child.* 17 (1), 111–127.
- Hirst, L.T., 1972. An Investigation of the Effects of Daily, Thirty-Minute Home Practice Sessions upon Reading Achievement with Second Year Elementary Pupils. Dissertation submitted to the University of Kentucky, Lexington, KY.
- Hitchcock, J., Dimino, J., Kurki, A., Wilkins, C., Gersten, R., 2011. The Impact of Collaborative Strategic Reading on the Reading Comprehension of Grade 5 Students in Linguistically Diverse Schools. U.S. Department of Education, National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, Washington, DC.
- Holmes, T., McConnell, B., 1990. Full-day versus half-day kindergarten: an experimental study. In: Paper Presented at the Annual Meeting of the Educational Research Association, Boston, MA.
- Holmstrom, B., Milgrom, P., 1991. Multitask principal-agent analyses: incentive contracts, asset ownership, and job design. *J. Law, Econ. Organ.* 7, 24–52.
- Hopkins, K., Bracht, G., 1975. Ten-year stability of verbal and nonverbal IQ scores. *Am. Educ. Res. J.* 12 (4), 469–477.
- Hotard, S., Cortez, M., 1983. Computer Assisted Instruction as an Enhancer of Remediation. The Title 1 Program, Lafayette Parish, LA.
- Hoxby, C., 2002. School Choice and School Productivity. NBER Working Paper no. 8873.

- Hoxby, C., Leigh, A., 2004. Pulled away or pushed out? explaining the decline of teacher aptitude in the United States. *Am. Econ. Rev.* 94 (2), 236–240.
- Hoxby, C., Murarka, S., 2009. Charter Schools in New York City: Who Enrolls and How They Affect Their Students' Achievement. NBER Working Paper no. 14852.
- Isenberg, E., Glazerman, S., Bleeker, M., Johnson, A., Lugo-Gil, J., Grider, M., Dolfin, S., Britton, E., Ali, M., 2009. Impacts of Comprehensive Teacher Induction: Results from the Second Year of a Randomized Controlled Study. U.S. Department of Education, National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, Washington, DC.
- Jackson, C., 2010. A Stitch in Time: The Effects of a Novel Incentive-based High-school Intervention on College Outcomes. NBER Working Paper no. w15722.
- Jacob, B., 2004. Remedial education and student achievement: a regression-discontinuity analysis. *Rev. Econ. Statistics* 86 (1), 226–244.
- James-Burdumy, S., Dynarski, M., Moore, M., Deke, J., Mansfield, W., Pistorino, C., Warner, E., 2005. When Schools Stay Open Late: The National Evaluation of the 21<sup>st</sup> Century Community Learning Centers Program: Final Report. U.S. Department of Education, National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, Washington, DC.
- James-Burdumy, S., Mansfield, W., Deke, J., Carey, N., Lugo-Gil, J., Hershey, A., Douglas, A., Gersten, R., Newman-Gonchar, R., Dimino, J., Faddis, B., Pendleton, A., 2008. Effectiveness of Selected Supplemental Reading Comprehension Interventions: Impacts on a First Cohort of Fifth-Grade Students. U.S. Department of Education, National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, Washington, DC.
- Jenkins, J., Peyton, J., Sanders, E., Vadasy, P., 2004. Effects of reading decodable texts in supplemental first-grade tutoring. *Sci. Stud. Read.* 8 (1), 53–85.
- Jeynes, W., 2005. Parental Involvement and Student Achievement: A Meta-Analysis. Harvard Family Research Project, Cambridge, MA.
- Jeynes, W., 2007. The relationship between parental involvement and urban secondary school student academic achievement: a meta-analysis. *Urban Educ.* 42 (1), 82–110.
- Jones, S., Brown, J., Aber, J., 2011. Two-year impacts of a universal school-based social-emotional and literacy intervention: an experiment in translational developmental research. *Child Dev.* 82 (2), 533–554.
- Joyce, B., Showers, B., 1988. Student Achievement through Staff Development. Longman, White Plains, NY.
- Kántor, Z., 2011. Ethnic or social integration? The Roma in Hungary. In: Kahane, M., Zimmerman, K. (Eds.), *Ethnic Diversity in European Labor Markets: Challenges and Solutions*. Edward Elgar Publishing, Cheltenham, UK, pp. 137–162.
- Kane, T., Staiger, D., 2008. Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation. NBER Working Paper no. 14607.
- Karper, J., Melnick, S., 1993. The effectiveness of team accelerated instruction on high achievers in mathematics. *J. Instr. Psychol.* 20 (1), 49–55.
- Kautz, T., Heckman, J., Diris, R., ter Weel, B., Borghans, L., 2014. Fostering and Measuring Skills: Improving Cognitive and Non-Cognitive Skills to Promote Lifetime Success. NBER Working Paper no. 20749.
- Kemple, J., Snipes, J., 2000. Career Academies: Impacts on Students' Engagement and Performance in High School. Manpower Demonstration Research Corporation.
- Kim, J., 2005. Project READS (Reading Enhances Achievement during Summer): Results from a Randomized Field Trial of a Voluntary Summer Reading Intervention. Paper presented at Princeton University, Education Research Section, Princeton, NJ.
- Kim, J., 2006. Effects of a voluntary summer reading intervention on reading achievement: results from a randomized field trial. *Educ. Eval. Policy Anal.* 28 (4), 335–355.
- Kim, J., 2007. The effects of a voluntary summer reading intervention on reading activities and reading achievement. *J. Educ. Psychol.* 99 (3), 505–515.
- Kim, J., Olson, C., Scarcella, R., Kramer, J., Pearson, M., van Dyk, D., Collins, P., Land, R., 2011. A randomized experiment of a cognitive strategies approach to text-based analytical writing for mainstreamed Latino English language learners in grades 6 to 12. *J. Res. Educ. Eff.* 4 (3), 231–263.

- Klaus, R., Gray, S., 1968. The Early Training Project for disadvantaged children: a report after five years. *Monogr. Soc. Res. Child Dev.* 33 (4) iii–iv+1–66.
- Klibanoff, L., Haggart, S., 1981. Summer Growth and the Effectiveness of Summer School. RMC Research Corporation, Mountainview, CA.
- Kling, J., Liebman, J., Katz, L., 2007. Experimental analysis of neighborhood effects. *Econometrica* 75 (1), 83–119.
- Knox, M., 1996. An Experimental Study of the Effects of The Accelerated Reading Program and a Teacher Directed Program on Reading Comprehension and Vocabulary of Fourth and Fifth Grade Students. Dissertation submitted to the University of South Florida.
- Knudson, E., Heckman, J., Cameron, J., Shonkoff, J., 2006. Economic, neurobiological, and behavioral perspectives on building America's future workforce. *Proc. Natl. Acad. Sci. U.S.A.* 103 (27), 10155–10162.
- Koch, E., 1965. Homework in arithmetic. *Arith. Teacher* 12 (1), 9–13.
- Kohen, D., Brooks-Gunn, J., Leventhal, T., Hertzman, C., 2002. Neighborhood income and physical and social disorder in Canada: associations with young children's competencies. *Child Dev.* 73 (6), 1844–1860.
- Konstantopoulos, S., Miller, S., van der Ploeg, A., 2013. The impact of Indiana's system of interim assessments on mathematics and reaching achievement. *Educ. Eval. Policy Anal.* 35 (4), 481–499.
- Kremer, M., Edward, M., Rebecca, T., 2009. Incentives to learn. *Rev. Econ. Statistics* 91 (3), 437–456.
- Krueger, A., 1999. Experimental estimates of education production functions. *Q.J. Econ.* 114 (2), 497–532.
- Kuhn, M., Schwanenflugel, P., Morris, R., Morrow, L., Woo, D., Meisinger, E., Sevcik, R., Bradley, B., Stahl, S., 2006. Teaching children to become fluent and automatic readers. *J. Lit. Res.* 38 (4), 357–387.
- Layton, L., August 04, 2015. Study: billions of dollars in annual teacher training is largely a waste. Wash. Post.
- Layzer, J., Layzer, C., Goodson, B., Price, C., 2007. Evaluation of Child Care Subsidy Strategies: Findings from Project Upgrade in Miami-Dade County. Abt Associates, Cambridge, MA.
- Lesnick, J., 2006. A Mixed-Method Multi-Level Randomized Evaluation of the Implementation and Impact of an Audio-Assisted Reading Program for Struggling Readers. Dissertation submitted to the University of Pennsylvania.
- Levitt, S., List, J., 2009. Field experiments in economics: the past, the present, and the future. *Eur. Econ. Rev.* 53 (1), 1–18.
- Lipsey, M., Wilson, D., 2000. Practical Meta Analysis. Sage Publications, Thousand Oaks, California.
- Loucks-Horsley, S., Mundry, S., Hewson, P., Love, N., Stiles, K., 1998. Designing Professional Development for Teachers of Mathematics and Science. Corwin Press, Thousand Oaks, CA.
- Ludwig, J., Duncan, G., Gennetian, L., Katz, L., Kessler, R., Kling, J., Sanbonmatsu, L., 2012. Neighborhood effects on the long-term well-being of low-income adults. *Science* 337 (6101), 1505–1510.
- Ludwig, J., Sanbonmatsu, L., Gennetian, L., Adam, E., Duncan, G., Katz, L., Kessler, R., Kling, J., Tessler Lindau, S., Whitaker, R., McDade, T., 2011. Neighborhoods, obesity, and Diabetes—A randomized social Experiment. *N. Engl. J. Med.* 365 (16), 1509–1519.
- Magnuson, K., Duncan, G., 2002. Parents in Poverty. In: Bornstein, M. (Ed.), *Handbook of Parenting*, second ed. Lawrence Erlbaum Associates, Mahwah, NJ.
- Malamud, O., Pop-Eleches, C., 2011. Home computer use and the development of human capital. *Q. J. Econ.* 126 (2), 987–1027.
- Manuel, S., 1987. The Relationship between Supplemental Computer Assisted Mathematics Instruction and Student Achievement. Dissertation submitted to the University of Nebraska — Lincoln.
- Marshall, H., Magruder, L., 1960. Relations between parent money education practices and Children's knowledge and use of money. *Child Dev.* 31 (2), 253–284.
- Mathes, P., Babyak, A., 2001. The effects of peer-assisted learning strategies for first-grade readers with and without additional mini-skills lessons. *Learn. Disabil. Res. Pract.* 16 (1), 28–44.
- Mathes, P., Denton, C., Fletcher, J., Anthony, J., Francis, D., Schatschneider, C., 2005. The effects of theoretically different instruction and student characteristics on the skills of struggling readers. *Read. Res. Q.* 40 (2), 148–182.
- May, H., Gray, A., Gillespie, J., Sirinides, P., Sam, C., Goldsworthy, H., Armijo, M., Tognatta, M., 2013. Evaluation of the I3 Scale-up of Reading Recovery. CPRE, Philadelphia, PA.

- Mayer, D., Peterson, P., Myers, D., Tuttle, C.C., Howell, W., 2002. School Choice in New York City after Three Years: An Evaluation of the School Choice Scholarships Program. Final Report. Mathematica Policy Research, Princeton, NJ.
- Mayfield, L., 2000. The effects of structured one-on-one tutoring in sight word recognition of first grade students at-risk for reading failure. Paper Presented at the Mid-South Educational Research Association Annual Meeting, Bowling Green, KY.
- Maynard, R., Murnane, R., 1979. The effects of a negative income tax on school performance: results of an experiment. *J. Hum. Resour.* 14 (4), 463–476.
- McCall, W., 1923. How to Experiment in Education. Macmillan, New York.
- McGill-Franzen, A., Allington, R., Yokoi, L., Brooks, G., 1999. Putting books in the classroom seems necessary but not sufficient. *J. Educ. Res.* 93 (2), 67–74.
- McLoyd, V., 1998. Socioeconomic disadvantage and child development. *Am. Psychol.* 53 (2), 185–204.
- Mehran, M., White, K., 1988. Parent tutoring as a supplement to compensatory education for first-grade children. *Remedial Spec. Educ.* 9 (3), 35–41.
- Meyer, E., Van Klaveren, C., 2013. The effectiveness of extended day programs: evidence from a randomized field experiment in the Netherlands. *Econ. Educ. Rev.* 36 (C), 1–11.
- Meyer, B., Wijekumar, K., Lin, Y., 2011. Individualizing a web-based structure strategy intervention for fifth graders' comprehension of nonfiction. *J. Educ. Psychol.* 103 (1), 140–168.
- Michalopoulos, C., Tattrie, D., Miller, C., Robins, P.K., Morris, P., Gyarmati, D., Redcross, C., Foley, K., Ford, R., 2002. Making Work Pay: Final Report on the Self-sufficiency Project for Long-term Welfare Recipients. Social Research and Demonstration Corporation, Ottawa, Canada.
- Miller, G., Jaciw, A., 2007. Comparative Effectiveness of Scott Foresman Science: A Report of Randomized Experiments in Five School Districts. Empirical Education Inc., Palo Alto, CA.
- Mischel, W., Ebbesen, E., Raskoff Zeiss, A., 1972. Cognitive and attentional mechanisms in delay of gratification. *J. Personality Soc. Psychol.* 21 (2), 204–218.
- Miyake, A., Kost-Smith, L., Finkelstein, N., Pollock, S., Cohen, G., Ito, T., 2010. Reducing the gender achievement gap in college science: a classroom study of values affirmation. *Science* 330 (60008), 1234–1237.
- Mooney, P., 2003. An Investigation of the Effects of a Comprehensive Reading Intervention on the Beginning Reading Skills of First Graders at Risk for Emotional and Behavioral Disorders. Dissertation submitted to the University of Nebraska – Lincoln.
- Morais de Sá e Silva, M., 2008. Opportunity NYC: A Performance-Based Conditional Cash Transfer Programme. International Poverty Centre Working Paper no. 49.
- Morrow, L., 1992. The impact of a literature-based program on literary achievement, use of literature, and attitudes of children from minority backgrounds. *Read. Res. Q.* 27 (3), 250–275.
- Morrow-Howell, N., Jonson-Reid, M., McCrary, S., Lee, Y., Spitznagel, E., 2009. Evaluation of Experience Corps: Student Reading Outcomes. Center for Social Development, George Warren Brown School of Social Work, Washington University in St. Louis, St. Louis, MO.
- Mosteller, F., Boruch, R., 2002. Evidence Matters: Randomized Trials in Education Research. Brookings Institution Press, Washington, DC.
- Moynihan, D., 1969. On Understanding Poverty: Perspectives from the Social Sciences. Basic Books, New York, NY.
- Muralidharan, K., Sundararaman, V., 2011. Teacher performance pay: experimental evidence from India. *J. Political Econ.* 119 (1), 39–77.
- Murnane, R., 1975. The Impact of School Resources on the Learning of Inner City Children. Ballinger Publishing, Cambridge, MA.
- National Alliance for Public Charter Schools, 2009. Public Charter Schools Dashboard, Charter School Market Share.
- National Alliance for Public Charter Schools, 2015. A Growing Movement: America's Largest Charter School Communities: Tenth Annual Edition. Washington, DC.
- National Telecommunication and Information Administration, 2011. Exploring the Digital Nation: Computer and Internet Use at Home. National Telecommunications and Information Administration, U.S. Department of Commerce, Washington, DC.

- Neal, D., Johnson, W., 1996. The role of premarket factors in black-white wage differences. *J. Political Econ.* 104 (5), 869–895.
- Neal, D., 2011. The Design of Performance Pay Systems in Education. NBER Working Paper no. 16710.
- Nelson, C., 2000. Neural plasticity and human development: the role of early experience in sculpting memory systems. *Dev. Sci.* 3 (2), 115–136.
- Nelson, R., 1959. An Experiment with class size in the teaching of elementary economics. *Educ. Rec.* 4, 241–275.
- Nelson, J., Stage, S., 2007. Fostering the Development of Vocabulary Knowledge and Reading Comprehension through Contextually-Based Multiple Meaning Vocabulary Instruction. Special Education and Communication Disorders Faculty Publications. Paper 28.
- Nelson-Royes, A., 2015. Why Tutoring?: A Way to Achieve Success in School. Rowman and Littlefield, Lanham, MD.
- Newman, D., Finney, P., Bell, S., Turner, H., Jaciw, A., Zacamy, J., Gould, L., Garcia, S., 2012. Evaluation of the Effectiveness of the Alabama Math, Science, and Technology Initiative (AMSTI). U.S. Department of Education, National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, Washington, DC.
- Newport, E., 1990. Maturational constraints on language learning. *Cognitive Sci.* 14 (11), 11–28.
- Nordin, M., Rooth, D.-O., 2007. Income Gap between Natives and Second Generation Immigrants in Sweden: Is Skill the Explanation? IZA Discussion Paper no. 2759.
- Nunnery, J., Ross, S., McDonald, A., 2006. A randomized experimental evaluation of the impact of Accelerated Reader/ Reading Renaissance implementation on reading achievement in grades 3 to 6. *JESPAR* 11 (1), 1–18.
- Nussbaum, S., 2010. The Effects of Brain Gym as a General Educational Intervention: Improving Academic Performance and Behaviors. Dissertation submitted to Northcentral University.
- Nye, C., Turner, H., Schwartz, J., 2006. Approaches to Parental Involvement for Improving the Academic Performance of Elementary School Children in Grades K-6. Harvard Family Research Project, Cambridge, MA.
- Olds, D., Henderson, C., Cole, R., 1998. Long-term effects of nurse home visitation on children's criminal and antisocial behavior: 15-year follow-up of a randomized controlled trial. *JAMA* 280, 1238–1244.
- Olds, D., Robinson, J., O'Brien, R., 2002. Home visiting by paraprofessionals and nurses: a randomized, controlled trial. *Pediatrics* 100, 486–496.
- O'Neill, June 1990. The role of human capital in earnings differences between black and white men. *J. Econ. Perspect.* 4 (4), 25–45.
- Oreopoulos, P., 2003. The long-run consequences of living in a poor neighborhood. *Q. J. Econ.* 118 (4), 1533–1575.
- Parsad, B., Lewis, L., Farris, E., 2001. Teacher Preparation and Professional Development: 2000. U.S. Department of Education, National Center for Education Statistics, Washington, DC.
- Parsons, C., Smeeding, T., 2008. Immigration and the Transformation of Europe. Cambridge University Press, Cambridge, UK.
- Peeples, R., 1996. The Impact of Parental Training in Methods to Aid Beginning Reading on Reading Achievement and Reading Attitudes of First-Grade Students. Dissertation submitted to the Florida State University.
- Phillips, L., Norris, S., Mason, J., Kerr, B., 1990. Effect of Early Literacy Intervention on Kindergarten Achievement. Technical Reports: Center for the Study of Reading, University of Illinois at Urbana-Champaign, Champaign, IL.
- Phillips, M., Brooks-Gunn, J., Duncan, G., Klebanov, P., Crane, J., 1998. Family background, parenting practices, and the black-white test score gap. In: Jenkins, C., Phillips, M. (Eds.), *The Black-White Test Score Gap*. Brookings Institution Press, Washington, DC, pp. 102–145.
- Pinkner, S., 1994. *The Language Instinct*. Harper Perennial Modern Classics, New York, NY.
- Pinnell, G., Lyons, C., DeFord, D., Bryk, A., Seltzer, N., 1994. Comparing instructional models for the literacy education of high risk first graders. *Read. Res. Q.* 29, 8–39.
- Porwell, P.J., 1978. Class Size: A Summary of Research. Educational Research Service, Arlington, VA.
- Potts, M., 1967. The effect of second-language instruction on the reading proficiency and general school achievement of primary-grade children. *Am. Educ. Res. J.* 4 (7), 367–373.

- Powell, D., Diamond, K., Burchinal, M., Koehler, M., 2010. Effects of an early literacy professional development intervention on Head Start teachers and children. *J. Educ. Psychol.* 102 (2), 299–312.
- Powell-Smith, K., Stoner, G., Shinn, M., Good III, R., 2000. Parent tutoring in reading using literature and curriculum materials: impact on student reading achievement. *Sch. Psychol. Rev.* 29 (1), 5–27.
- Preschool Curriculum Evaluation Research Consortium, 2008. Effects of Preschool Curriculum Programs on School Readiness: Report from the Preschool Curriculum Evaluation Research Initiative. U.S. Department of Education, National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, Washington, DC.
- Pullen, P., Lane, H., Monaghan, M., 2004. Effects of a volunteer tutoring model on the early literacy development of struggling first grade students. *Read. Res. Instruct.* 43 (4), 21–40.
- Puma, M., Bell, S., Cook, R., Heid, C., 2010. Head Start Impact Study Final Report. U.S. Department of Health and Human Services, Administration for Children and Families, Washington, DC.
- Ramey, C., Campbell, F., 1991. Poverty, early childhood education, and academic competence: the Abecedarian experiment. In: Houston, A. (Ed.), *Children Reared in Poverty*. Cambridge University Press, New York, NY.
- Ramey, S., Ramey, C., Phillips, M., Lanzi, R., Brezausek, C., Katholi, C., Snyder, S., 2000. Head Start Children's Entry into Public School: A Report on the National Head Start/Public School Early Childhood Transition Demonstration Study. Civitan International Research Center, The University of Alabama at Birmingham, Birmingham, AL.
- Randel, B., Beesley, A., Apthorp, H., Clark, T., Wang, X., Cicchinelli, L., Williams, J., 2011. Classroom Assessment for Student Learning: Impact on Elementary School Mathematics in the Central Region. National Center for Educational Evaluation and Regional Assistance, Washington, DC.
- Raver, C., Jones, S., Li-Grining, C., Zhai, F., Bub, K., Pressler, E., 2011. CSPR's impact on low-income preschoolers' preacademic skills: self-regulation as a mediating mechanism. *Child Dev.* 82 (1), 362–378.
- Resendez, M., Azin, M., 2012. A Study on the Effects of Houghton Mifflin Harcourt's "Journeys" Program: Year 1 Final Report. Planning, Research and Evaluation Services Associates, Inc.
- Riccio, J., Dechausay, N., Greenberg, D., Miller, C., Rucks, Z., Verma, N., 2010. Toward Reduced Poverty across Generations: Early Findings from New York City's Conditional Cash Transfer Program. MDRC.
- Riccio, J., Dechausay, N., Miller, C., Nuñez, S., Verma, N., Yang, E., 2013. Conditional Cash Transfers in New York City: The Continuing Story of the Opportunity NYC-Family Rewards Demonstration (New York: MDRC).
- Rickford, J., 1999. African American Vernacular English: Features, Evolution, Educational Implication. Blackwell Publishers, Malden, MA.
- Rivkin, S., Hanushek, E., Kain, J., 2005. Teachers, schools, and academic achievement. *Econometrica* 73 (2), 417–458.
- Rockoff, J., 2004. The impact of individual teachers on student achievement: evidence from panel data. *Am. Econ. Rev.* 94 (2), 247–252.
- Rockoff, J., Staiger, D., Kane, T., Taylor, E., 2012. Information and employee evaluation: evidence from a randomized intervention in public schools. *Am. Econ. Rev.* 102 (7), 3184–3213.
- Rodríguez-Planas, N., 2012. Longer-term impacts of mentoring, educational services, and learning incentives: evidence from a randomized trial in the United States. *Am. Econ. Rev.* 4 (4), 121–139.
- Rosenbaum, J., 1995. Changing the geography of opportunity by expanding residential choice: lessons from the gautreaux program. *Hous. Policy Debate* 6 (1), 231–269.
- Rothstein, J., von Wachter, T., 2017. Social experiments in the labor market. In: Duflo, E., Banerjee, A. (Eds.), *Handbook of Field Experiments*, vol. 2, pp. 95–322.
- Rouse, C.E., 1998. Private school vouchers and student achievement: an evaluation of the Milwaukee parental choice program. *Q. J. Econ.* 113 (2), 553–602.
- Rutherford, T., Kibrick, M., Burchinal, M., Richland, L., Conley, A., Osborne, K., Schneider, S., Duran, L., Coulson, A., Antenore, F., Daniels, A., Martinez, M., 2010. Spatial temporal mathematics at scale: an innovative and fully developed paradigm to boost math achievement among all learners. Paper presented at the Annual Convention of the American Educational Research Association, Denver, CO.

- Ryan, E.McI., 1964. A Comparative Study of the Reading Achievement of Second Grade Pupils in Programs Characterized by a Contrasting Degree of Parent Participation. Dissertation submitted to the School of Education, Indiana University, Bloomington, IN.
- Ryan, R., 1982. Control and information in the intrapersonal sphere: an extension of cognitive evaluation theory. *J. Personality Soc. Psychol.* 63, 397–427.
- Salaway, J., 2008. Efficacy of a Direct Instruction Approach to Promote Early Learning. Dissertation submitted to Duquesne University.
- Sallis, J., McKenzie, T., Kolody, B., Lewis, M., Marshall, S., Rosengard, P., 1999. Effects of health-related physical education on academic achievement: project SPARK. *Res. Q. Exerc. Sport* 70 (2), 127–134.
- Sanbonmatsu, L., Kling, J., Duncan, G., Brooks-Gunn, J., 2006. Neighborhoods and academic achievement: results from the Moving to Opportunity experiment. *J. Hum. Resour.* 41 (4), 649–691.
- Sanbonmatsu, L., Ludwig, J., Katz, L., Gennetian, L., Duncan, G., Kessler, R., Adam, E., McDade, T., Lindau, S.T., 2011. Moving to Opportunity for Fair Housing Demonstration Program: Final Impacts Evaluation. U.S. Department of Housing and Urban Development, Washington, DC.
- Schmitt, J., Wadsworth, J., 2006. Changing Patterns in the Relative Economic Performance of Immigrants to Great Britain and the U.S., 1980–2000. CEPR, Cambridge, MA.
- Schultz, T.P., 2000. Final Report: The Impact of PROGRESA on School Enrollments. International Food Policy Research Institute, Washington, DC.
- Schwartz, R., 2005. Literacy learning of at-risk first-grade students in the reading recovery early intervention. *J. Educ. Psychol.* 97 (2), 257–267.
- Schweinhart, L., Barnes, H., Weikart, D.P., 1993. Significant Benefits: The HighScope Perry Preschool Study through Age 27. HighScope Press, Ypsilanti, MI.
- Schweinhart, L., Montie, J., Xiang, Z., Barnett, W., Belfield, C., Nores, M., 2005. Lifetime Effects: The High/Scope Perry Preschool Study through Age 40. HighScope Press, Ypsilanti, MI.
- Sivin-Kachala, J., Bialo, E., 2005. Evaluation Research on the Effectiveness of Fluency Formula: Final Report. Interactive Educational Systems Design, Inc.
- Skoufias, E., 2005. PROGRESA and its Impacts on the Welfare of Rural Households in Mexico. International Food Policy Research Institute, Washington, DC.
- Slavin, R., 2010. Can financial incentives enhance educational outcomes? Evidence from international experiments. *Educ. Res. Rev.* 5 (1), 68–80.
- Slavin, R., Karweit, N., 1985. Effects of whole class, ability grouped, and individualized instruction on mathematics achievement. *Am. Educ. Res. J.* 22 (3), 351–367.
- Slavin, R., Leavey, M., Madden, N., 1984. Combining cooperative learning and individualized instruction: effects on student mathematics achievement, attitudes, and behaviors. *Elem. Sch. J.* 84 (4), 409–422.
- Slavin, R., Madden, N., Calderón, M., Chamberlain, A., Hennessy, M., 2011. Reading and language outcomes of a multiyear randomized evaluation of transitional bilingual education. *Educ. Eval. Policy Anal.* 33 (1), 47–58.
- Sloan, H., 1993. Direct Instruction in Fourth and Fifth-Grade Classrooms. Dissertation submitted to Purdue University.
- Snider, V., Crawford, D., 1996. Action research: implementing Connecting Math Concepts. *Eff. Sch. Pract.* 15 (2), 17–26.
- Somers, M., Corrin, W., Sepanik, S., Salinger, T., Levin, J., Zmach, C., Wong, E., Strasberg, P., Silverberg, M., 2010. The Enhanced Reading Opportunities Study: Final Report. U.S. Department of Education, National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, Washington, DC.
- Spörer, N., Brunstein, J., Kieschke, U., 2009. Improving students' reading comprehension skills: effects of strategy instruction and reciprocal teaching. *Learn. Instr.* 19 (3), 272–286.
- Springer, M., Ballou, D., Hamilton, L., Vi-Nhuan, L., Lockwood, J.R., McCaffrey, D.F., Pepper, M., Stecher, B.M., 2010. Teacher Pay for Performance. NCPI, Nashville, TN.
- Springer, M., Pane, J., Vi-Nhuan, L., McCaffrey, D.F., Burns, S., Hamilton, L., Stecher, B.M., 2012. Team pay for performance. *Educ. Eval. Policy Anal.* 34 (4), 367–390.
- St Pierre, R., Layzer, J., Goodson, B., Bernstein, L., 1997. National Impact Evaluation of the Comprehensive Child Development Program. Abt Associates, Cambridge, MA.

- Sumi, W., Woodbridge, M., Javitz, H., Thornton, P., Wagner, M., Rouspil, K., Yu, J., Seeley, J., Walker, H., Golly, A., Small, J., Feil, E., Severson, H., 2012. Assessing the effectiveness of first step to Success: are short-term results the first step to long-term behavioral improvements? *J. Emot. Behav. Disord.* 21 (1), 1–14.
- Taylor, E., Tyler, J., 2012. The effect of evaluation on teacher performance. *Am. Econ. Rev.* 102 (7), 3628–3651.
- Tharp, R., 1982. The effective instruction of comprehension: results and description of the Kamehameha Early Education Program. *Read. Res. Q.* 17 (4), 503–527.
- The New Teacher Project (TNTP), 2015. *The Mirage: Confronting the Hard Truth about Our Quest for Teacher Development*. The New Teacher Project, Brooklyn, NY.
- Tizard, J., Schofield, W., Hewison, J., 1982. Collaboration between teachers and parents in assisting children's reading. *Educ. Psychol.* 52 (1), 1–15.
- Todd, P., Wolpin, K., 2003. On the specification and estimation of the production function for cognitive achievement. *Econ. J.* 113 (485), F3–F33.
- Tolan, P., Gorman-Smith, D., Henry, D., 2004. Supporting families in a high-risk setting: proximal effects of the SAFEChildren preventive intervention. *J. Consult. Clin. Psychol.* 72 (5), 855–869.
- Torgesen, J., Wagner, R., Rashotte, C., 1997. Prevention and remediation of severe reading disabilities: keeping the end in mind. *Sci. Stud. Read.* 1 (3), 217–234.
- Torgesen, J., Myers, D., Schirm, A., Stuart, E., Vartivarian, S., Mansfield, W., Stancavage, F., Durno, D., Javorsky, R., Haan, C., 2006. National Assessment of Title 1: Interim Report: Volume II: Closing the Reading Gap. U.S. Department of Education, National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, Washington, DC.
- Torgesen, J., Wagner, R., Rashotte, C., Herron, J., Lindamood, P., 2010. Computer-assisted instruction to prevent early reading difficulties in students at risk for dyslexia: outcomes from two instructional approaches. *Ann. Dyslexia* 60 (1), 40–56.
- Tucker, M., 2011. Teacher quality: what's wrong with U.S. strategy? *Educ. Leadersh.* 49 (4), 42–46.
- Turner, L., 1985. An Evaluation of the Effects of Paired Learning in a Mathematics Computer-Assisted Instruction Program. Dissertation submitted to Arizona State University.
- Tuttle, C., Gill, B., Gleason, P., Knechtel, V., Nichols-Barrer, I., Resch, A., 2013. *KIPP Middle Schools: Impacts on Achievement and Other Outcomes*. Final Report. Mathematica Policy Research, Princeton, NJ.
- U.S. Department of Education, 2009. *State and Local Implementation of the No Child Left Behind Act*. U.S. Department of Education, Washington, DC.
- U.S. Department of Education, 2014. *Fiscal Year 2015 Education Budget Summary and Background Information*. U.S. Department of Education, Washington, DC.
- U.S. Department of Education, 2015. *Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, what Works Clearinghouse*.
- U.S. Government Accountability Office, 2014. *K-12 Education: Characteristics of the Investing in Innovation Fund*. U.S. Government Accountability Office, Washington, DC.
- Uhry, J., Shepherd, M., 1993. Segmentation/spelling instruction as part of a first-grade reading program: effects on several measures of reading. *Read. Res. Q.* 28 (3), 218–233.
- United Nations Development Programme, 2010. *Human Development Report* (New York: United Nations).
- Vadasy, P., Sanders, E., 2008. Repeated reading instruction: outcomes and interactions with readers' skills and classroom instruction. *J. Educ. Psychol.* 100 (2), 272–290.
- Vadasy, P., Sanders, E., Tudor, S., 2007. Effectiveness of paraeducator-supplemented individual instruction: beyond basic decoding skills. *J. Learn. Disabil.* 40 (6), 508–525.
- Vaughn, S., Klingner, J., Swanson, E., Boardman, A., Roberts, G., Mohammed, S., Stillman-Spisak, S., 2011. Efficacy of collaborative strategic reading with middle school students. *Am. Educ. Res. J.* 48 (4), 938–964.
- Vaughn, S., Swanson, E., Roberts, G., Wanzek, J., Stillman-Spisak, S., Solis, M., Simmons, D., 2013. Improving reading comprehension and social studies knowledge in middle school. *Read. Res. Q.* 48 (1), 77–93.

- Vigdor, J., Ladd, H., 2010. Scaling the Digital Divide: Home Computer Technology and Student Achievement. NBER Working Paper no. 16078.
- von, L., Dietrich, H., 2011. Social and labor market integration of ethnic minorities in Germany. In: Kahanec, M., Zimmerman, K. (Eds.), *Ethnic Diversity in European Labor Markets: Challenges and Solutions*. Edward Elgar Publishing, Cheltenham, UK, pp. 109–136.
- Wang, C., Algozzine, B., 2008. Effects of targeted intervention on early literacy skills of at-risk students. *J. Res. Child. Educ.* 22 (4), 425–439.
- Warren, P., 2009. The Effects of Training Parents in Teaching Phonemic Awareness on the Phonemic Awareness and Early Reading of Struggling Readers. Dissertation submitted to Auburn University.
- Wasik, B., Bond, M., 2001. Beyond the pages of a book: interactive book reading and language development in preschool classrooms. *J. Educ. Psychol.* 93 (2), 243–250.
- Wasik, B., Bond, M., Hindman, A., 2006. The effects of a language and literacy intervention on Head Start children and teachers. *J. Educ. Psychol.* 98 (1), 63–74.
- Weikart, D., Deloria, D., Lawser, S., Wiegerink, R., 1970. Longitudinal Results of the Ypsilanti Perry Preschool Project. In: Ypsilanti, M.I. (Ed.). *High/Scope Educational Research Foundation*.
- West, M., Peterson, P., Campbell, D., 2001. School Choice in Dayton, Ohio after Two Years: An Evaluation of the Parents Advancing Choice in Education Program. KSG Working Paper No. RWP02–021.
- Wheldall, K., 2000. Does Rainbow Reading add value to an intensive literacy intervention program for low-progress readers? An experimental evaluation. *Educ. Rev.* 52 (1), 29–36.
- Wijekumar, K., Hitchcock, J., Turner, H., Lei, P., Peck, K., Park, O., 2009. A Multisite Cluster Randomized Trial of the Effects of Compass Learning Odyssey Math on the Math Achievement of Selected Grade 4 Students in the Mid-Atlantic Region. U.S. Department of Education, National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, Washington, DC.
- Wijekumar, K., Meyer, B., Lei, P., 2012. Large-scale randomized controlled trial with 4<sup>th</sup> graders using intelligent tutoring of the structure strategy to improve nonfiction reading comprehension. *Educ. Technol. Res. Dev.* 60 (6), 987–1013.
- Wilkerson, S., Shannon, L., Herman, T., 2006. An Efficacy Study on Scott Foresman's Reading Street Program: One Year Report. Magnolia Consulting, Lousia, VA.
- Wilson, T., Linville, P., 1982. Improving the academic performance of college freshmen: attribution therapy revisited. *J. Psychol. Soc. Psychol.* 42 (2), 367–376.
- Winship, S., Owen, S., 2013. The Brookings Social Genome Model. Brookings, Washington, DC.
- Witte, J., 1997. Achievement effects of the Milwaukee voucher program. In: Paper Presented at the 1997 American Economics Association Annual Meeting, New Orleans, LA.
- Witte, J., Sterr, T., Thorn, C., 1995. Fifth-year Report Milwaukee Parental Choice Program. LaFollette School Working Paper no. 1995–001.
- Wolf, P., Gutmann, B., Puma, M., Kisida, B., Rizzo, L., Elissa, N., Carr, M., 2010. Evaluation of the DC Opportunity Scholarship Program. National Center for Education Evaluation and Regional Assistance, Washington, DC.
- Worrall, J., 2007. Evidence in medicine and evidence-based medicine. *Philos. Compass* 2 (6), 981–1022.
- Yeager, D., Walton, G., 2011. Social-psychological interventions in education. *Rev. Educ. Res.* 81 (2), 267–301.
- Yeates, K., MacPhee, D., Campbell, F., Ramey, C., 1983. Maternal IQ and home environment as determinants of early childhood intellectual competence: a developmental analysis. *Dev. Psychol.* 19, 731–739.
- York, B., Loeb, S., 2014. One Step at a Time: The Effects of an Early Literacy Text Messaging Program for Parents and Preschoolers. NBER Working Paper no. 20659.
- Ysseldyke, J., Bolt, D., 2007. Effect of technology-enhanced continuous progress monitoring on math achievement. *Sch. Psychol. Rev.* 36 (3), 453–467.
- Zvoch, K., Stevens, J., 2012. Summer school effects in a randomized field trial. *Early Child. Res. Q.* 28 (1), 24–32.

## CHAPTER 3

# Field Experiments in Education in Developing Countries

K. Muralidharan\*, §, ¶, a

\*University of California, San Diego, La Jolla, CA, United States

§NBER (National Bureau of Economic Research), Cambridge, MA, United States

¶Jameel Poverty Action Lab, Cambridge, MA, United States

E-mail: kamurali@ucsd.edu

## Contents

1. Introduction	324
2. Field Experiments in Education—A Short Overview	325
2.1 Background	325
2.2 The main research questions	326
2.3 The value of experiments in education and reasons for their growth	328
2.4 An alternative framing of the questions of interest	329
3. Selected Overview of Field Experiments in Education in Developing Countries	330
3.1 Demand-side interventions	330
3.2 School and student inputs	334
3.3 Pedagogy	338
3.4 Governance	343
3.5 Summary of evidence	348
4. Limitations of Field Experiments and Strategies for Mitigating Them	349
4.1 Production function versus policy parameters	349
4.2 Interpreting zero effects	351
4.3 External validity	353
4.3.1 <i>External validity in the same context: representativeness of study samples</i>	353
4.3.2 <i>External validity in the same context: implementer heterogeneity</i>	353
4.3.3 <i>External validity in the same context: varying intervention details</i>	354
4.3.4 <i>External validity in the same context: political economy</i>	355
4.3.5 <i>External validity concerns across contexts</i>	356
5. Conducting Field Experiments in Education in Developing Countries	357
5.1 Design	357
5.1.1 <i>Intervention design: what to look for?</i>	358
5.1.2 <i>Intervention design: what to avoid?</i>	360
5.1.3 <i>Experiment design: unit of randomization</i>	362
5.1.4 <i>Experiment design: power and baselines</i>	364
5.1.5 <i>Experiment design: cross-cutting designs and interactions</i>	366
5.2 Sampling, randomization, and implementation	367

<sup>a</sup> I thank Abhijit Banerjee, Alejandro Ganimian, Asim Khwaja, and Abhijeet Singh for several useful comments and discussions. All views represented here are my own, and not of any organization I am affiliated with.

5.2.1 Sampling and representativeness	367
5.2.2 Randomization	368
5.2.3 Implementation and follow-up	368
<b>5.3 Data collection</b>	<b>369</b>
5.3.1 Outcomes	369
5.3.2 Intermediate inputs, processes, and mechanisms	370
5.3.3 Long-term follow-up	371
<b>5.4 Analysis</b>	<b>372</b>
5.4.1 Main treatment effects	372
5.4.2 Heterogeneity	374
5.4.3 Mechanisms	376
5.4.4 Cost effectiveness	376
<b>6. Conclusion</b>	<b>376</b>
References	379

## Abstract

The study of education in developing countries has been transformed by the rapid increase in the feasibility and prevalence of field experiments over the past 15 years. This paper comprises three main sections. First, it illustrates the very broad range of research questions regarding education in developing countries that have been addressed using field experiments, and summarizes the most important patterns of findings from this body of research. Second, it discusses some of the limitations of field experiments and strategies for mitigating them through better design. Third, it provides a practical toolkit on design, implementation, measurement and data collection, analysis, and interpretation of field experiments in education. The main goal for this chapter is to serve as a reference for students, researchers, and practitioners by summarizing lessons learned, highlighting key open questions for future research, and providing guidance on how to design and implement high-quality field experiments in education in a way that maximizes what we learn from them.

## Keywords

Development; Education; Experimental design; Field experiments; Synthesis; Toolkit

## JEL Codes

C93; H42; I21; I25; I28; O15

## 1. INTRODUCTION

Perhaps no field in development economics in the past decade has benefited as much from the use of experimental methods as the economics of education. The rapid growth in high-quality studies on education in developing countries (many of which use randomized experiments) is perhaps best highlighted by noting that there have been *several* systematic reviews of this evidence aiming to synthesize findings for research and policy in *just the past three years*. These include [Muralidharan \(2013\)](#) (focused on India), [Glewwe et al. \(2014\)](#) (focused on school inputs), [Kremer et al. \(2013\)](#), [Krishnaratne et al. \(2013\)](#),

Conn (2014) (focused on sub-Saharan Africa), McEwan (2014), Ganimian and Murnane (2016), Evans and Popova (2015), Snistveit et al. (2016), and Glewwe and Muralidharan (2016). While these are not all restricted to experimental studies, they typically provide greater weight to evidence from randomized controlled trials (RCTs).

The reviews above are mostly written for policy audiences and aim to summarize the policy implications of the research on education in developing countries. In contrast, this chapter is mainly written for graduate students and young researchers aiming to conduct field experiments in education in developing countries. The chapter aims to achieve two goals. The first is to illustrate the broad range of studies that have been conducted in this area by providing a curated summary of the main insights from the research over the past 15 years. The second (and more important) goal is to provide a practical toolkit on design, implementation, measurement and data collection, analysis, and interpretation of field experiments in education. The chapter aims to serve as a reference for students, researchers, and practitioners to guide the design and implementation of high-quality field experiments in education in a way that maximizes what we learn from them.

The chapter is organized as follows. Section 2 provides a conceptual overview of the core research questions in education in developing countries, and the role of field experiments in answering them. Section 3 highlights some of the key research questions in education in developing countries that have been addressed by field experiments, and aims to synthesize the main insights that have been obtained from this research in the past 15 years. Section 4 discusses some of the important limitations of field experiments and ways in which they may be mitigated. Section 5 presents an extensive set of guidelines for students and practitioners on the effective design, implementation, data collection, analysis, and interpretation of field experiments. Section 6 concludes and discusses areas for future research.

## 2. FIELD EXPERIMENTS IN EDUCATION—A SHORT OVERVIEW

### 2.1 Background

Education and human capital are widely considered to be essential inputs for both aggregate growth and development (Lucas, 1990; Barro, 1991; Mankiw et al., 1992), as well as for enhancing the capabilities and freedoms of individuals, and thereby enabling them to contribute to and participate in the process of economic development (Sen, 1993). Thus, improving education outcomes in developing countries has been an important policy priority for both national governments as well as for the international development and donor community. The importance of education in the development agenda is perhaps best reflected by the fact that two of the eight United Nations Millennium Development Goals (MDGs) pertained to education (achieving universal primary education, and achieving gender parity at all levels of education—both by 2015). The post-2015

Sustainable Development Goals (SDGs) continue to prioritize inclusive and equitable quality education for all (goal 4).

Further, education (especially school education) in most countries is typically provided publicly by the government and financed by taxpayer contributions. It is beyond the scope of this chapter to analyze *why* this is the case, but there are at least three reasons for the preponderance of publicly financed and provided education. First, there is considerable evidence to suggest that a variety of supply- and demand-side constraints may prevent optimal education investments by households and optimal provision by markets, and that outcomes can be improved by a social planner (see [Glewwe and Muralidharan, 2016](#) for a review). Second, it is widely believed that education generates positive spillovers beyond the returns that accrue to individuals, which would also suggest an active role for governments in education financing and production.<sup>1</sup> Finally, an important noneconomic reason for publicly provided education may be states' desire to control curriculum and the content of education, which affect the formation of preferences and civic identity ([Kremer and Sarychev, 2008](#)).

Thus, financing and producing education is an important policy priority for most countries, and public spending on education is typically one of the three largest components of government budgets (the other two being defense and healthcare). However, while this is true for most countries, developing countries face especially acute challenges in achieving universal quality education. They have lower levels of school enrollment and completion and much poorer learning outcomes, and also have fewer public resources to spend on education. Thus, spending scarce public funds effectively is especially important in developing countries, where the opportunity cost of poor spending is higher.<sup>2</sup> As a result, a major area of focus for research on education in developing countries has been to understand the effectiveness (both absolute and relative) of various policy options to improve education outcomes.

## 2.2 The main research questions

Most experimental research on education in developing countries in the past decade has focused on two main policy questions. First, how should we increase school enrollment

<sup>1</sup> Models with complementarity in worker human capital in production (such as [Kremer, 1993](#)) predict spillovers. [Lucas \(1988\)](#) argues that human capital spillovers may be large enough to explain most of the long-run differences in per capita income between high- and low-income countries. [Moretti \(2004\)](#) provides evidence of such spillovers in the context of US-manufacturing workers and plants. One direct channel of spillovers for which there is evidence in a developing country context is that education promotes technology adoption, and that nonadopters (who may be less educated) learn from adopters, which is a positive spillover from education that would not have been accounted for in the decision making of individually optimizing agents ([Foster and Rosenzweig, 1995](#)).

<sup>2</sup> In principle, governments in developing countries should be able to borrow to undertake any investment where the social rate of return is greater than the cost of borrowing. In practice, financial markets typically constrain the extent of government borrowing, which places a hard budget constraint on public expenditure.

and attendance, and second, how should we improve learning outcomes? The two questions are closely related because increased enrollment and attendance are likely to be necessary preconditions for improving learning outcomes. Nevertheless, it is useful to think about the two problems distinctly because the school attendance decision is typically made by parents, whereas the extent to which increased school participation translates into improved learning outcomes is more likely to be affected by school-level factors.

For school participation and attendance, a simple model of optimizing households in the tradition of Becker (1962) and Ben-Porath (1967) yields the result that households will only invest in an additional year of education for their children if the present discounted value of the expected increase in benefits exceeds the costs of doing so. Thus, policies that seek to improve school participation typically aim to increase the immediate benefits to households of sending their children to school or to reduce the costs of doing so. The magnitude of the impact of these policies will in turn depend on the distribution of the household-child-specific unobservables that determine whether a given child enrolls in or attends school, and the extent to which the policy helps make it more attractive to do so.

For quality of learning, a standard education production function with certain additional assumptions (see Todd and Wolpin (2003) for a detailed exposition of these assumptions) allows the lagged test score to be treated as a sufficient statistic for representing prior inputs into learning, and for the use of a value-added model (VAM) to study the impact of changing contemporaneous inputs into education on test scores. Specifically, the typical VAM takes the form:

$$T_{i,t} = \gamma T_{i,t-1} + \beta \mathbf{X}_{i,t} + \varepsilon_{i,t} \quad (1)$$

where  $T_{i,t}$  represents test scores of child  $i$  at time  $t$ ,  $T_{i,t-1}$  represents the lagged test score, and  $\mathbf{X}_{i,t}$  represents a full vector of contemporaneous home ( $\mathbf{H}_{i,t}$ ) and school ( $\mathbf{S}_{i,t}$ ) inputs. While the production function above is linear in  $\mathbf{X}_{i,t}$  and is typically estimated this way, the specification does not have to be as restrictive, because  $\mathbf{X}_{i,t}$  can include nonlinear terms in individual inputs, as well as interaction terms between specific sets of inputs.

Given the budget constraints in public education, and the almost unlimited set of ideas for inputs and interventions that may improve education outcomes, an optimal policy approach to allocating scarce resources across the set of potential inputs would be to estimate the marginal return to providing a specific input and to compare it with the marginal cost of doing so (since these inputs are typically provided publicly) and to prioritize investments in diminishing order of the estimated return per dollar spent. Such an approach may also provide a basis for improving the effectiveness of education spending by pivoting existing expenditure away from less to more cost-effective expenditure items.

Since cost data are relatively easier to obtain,<sup>3</sup> the main practical challenge is one of estimating the marginal returns to different inputs. The economics of education literature has correspondingly devoted a lot of attention towards doing this and produced a large body of papers across several developing countries trying to estimate these returns for various inputs (see [Glewwe et al., 2014](#); [Glewwe and Muralidharan, 2016](#) for a review).

### 2.3 The value of experiments in education and reasons for their growth

The main challenge for nonexperimental studies (that use observational data) is the concern that variation in the specific input being studied ( $X_{i,t}$ ) is correlated with the unobserved error term ( $\varepsilon_{i,t}$ ), yielding biased estimates of  $\beta$ . In practice, this is quite likely to be true. For instance, communities and parents that care more about education are likely to be able to successfully lobby for more school inputs, and are also likely to provide unmeasured inputs into their children's education, which would lead to an upward bias on  $\beta$  estimated in cross-sectional data. In other cases, governments may target inputs to disadvantaged areas to improve equity in which case areas with increases in  $X_{i,t}$  may be negatively correlated with  $\varepsilon_{i,t}$ , yielding downward-biased estimates of  $\beta$ .

Thus, the value of experimental evaluations in this setting is that random assignment of the input (intervention) of interest solves this identification problem by ensuring that variation in  $X_{i,t}$  is orthogonal to variation in  $\varepsilon_{i,t}$ , thereby yielding unbiased estimates of  $\beta$  (with some caveats as noted in [Section 4](#)).<sup>4</sup> The importance of accounting for omitted variable bias in the evaluation of education interventions is starkly illustrated by [Glewwe et al. \(2004\)](#) who compare retrospective and prospective studies of the impact of classroom flipcharts on learning outcomes. When they use observational data, they find that flipcharts in classrooms appear to raise student test scores by  $0.2\sigma$ . However, when they conduct an RCT of flipcharts in classrooms, they find no impact on test scores at all, suggesting that the nonexperimental estimates were significantly biased upwards (even after controlling for observable factors). These results underscore the value of field experiments for program evaluation in developing countries and [Glewwe et al. \(2004\)](#) can be considered analogous to [LaLonde \(1986\)](#) in the US program evaluation literature, which showed that nonexperimental methods were not able to replicate the estimates from experimental evaluations of the impact of job-training programs.

While field experiments have improved causal inference in most topics in applied microeconomics, they have been particularly prevalent in the economics of education (especially in developing countries) in recent years. There are several reasons for this. First,

<sup>3</sup> In practice, even obtaining cost estimates of specific interventions is nontrivial (especially when it involves aggregating expenditure across multiple levels of government), but in principle, they can be reconstructed from government budget documents.

<sup>4</sup> See the companion chapter "The Econometrics of Randomized Experiments" by [Athey and Imbens \(2017\)](#) for more details.

many interventions in education are “modular” and therefore feasible to randomize at the student, classroom, or school level. Second, the outcome variables are quite well-defined and there is considerable agreement among economists on the key outcomes that programs should aim to improve (enrollment, attendance, and test scores<sup>5</sup>). Third, the large number of nonprofit organizations that work on education has made it feasible for researchers to find implementation partners who can design and deploy promising interventions to study. Fourth, since nongovernment implementation partners typically cannot (and are not expected to) work “everywhere”, it is politically and practically feasible for them to use a lottery to determine where their programs will be rolled out first, since this ensures fairness in program access in addition to enabling experimental evaluations of impact. Finally, data from regularly conducted nationwide surveys like ASER in India and Uwezo in East Africa show that despite considerable increases in school enrollment in developing countries, learning outcomes in these settings are very low (with a large fraction of students not being functionally literate or numerate at the end of primary school). The wide dissemination of these results has increased the demand from policy makers and funders of education programs for evidence on the impact of the programs they are funding, and for cost-effective ways of improving learning.

This confluence of factors has led to several high-quality experimental studies in education in developing countries that have both contributed evidence on the effectiveness of specific programs and also promoted a deeper understanding of the barriers to improving education outcomes in these settings. These include studies on interventions to improve parent-and-student demand for education, school and household inputs, classroom organization and pedagogy, and school governance and teacher effort (with some interventions combining features across this broad classification). When put together (with some caveats), this body of research also enables comparison of marginal costs and benefits across different kinds of education spending and can guide policy priorities over allocation of limited resources.

## 2.4 An alternative framing of the questions of interest

The use of experiments above has been motivated by wanting to measure the “impacts” of interventions and to estimate their cost effectiveness at improving school participation and learning outcomes. Yet, an alternative way of framing the question of interest is to ask: “What are the determinants of school participation and learning outcomes?” This approach focuses on understanding household decision-making regarding human capital investments as a function of household beliefs and preferences, and constraints including production function, budget, credit, and information constraints. Randomized

<sup>5</sup> While test scores are not the ultimate outcomes that a social planner cares about, studies using long-term longitudinal data find that interventions that raise test scores in school (such as having a better teacher in primary and middle school) also lead to better long-term labor market outcomes (Chetty et al., 2011).

evaluations of interventions are then interpreted less through the lens of their “impact” on outcomes of policy interest, and more through the lens of providing exogenous variation in the constraints above, which enables the researcher to better understand the determinants of school participation and learning outcomes (see [Attanasio, 2015](#) for an illustration of such an approach).

One way of thinking about the difference between these two ways of framing the question is that the approach in this section is more that of a “scientist” trying to understand the world, whereas the approach in [Section 2.3](#) is more that of an “engineer” trying to improve outcomes and solve problems (see [Mankiw, 2006](#) for a discussion of a similar distinction in approaches to macroeconomics). The synthesis of evidence in [Section 3](#) follows the approach in [Section 2.3](#) because most of the experimental research in education in recent years has been motivated by policy questions of how best to improve education outcomes in specific settings (this is also the style taken by the other systematic reviews referenced in the introduction). At the same time, there are important complementarities between the two approaches, and I argue in [Section 5](#) that studies that bridge this divide effectively will typically generate more generalizable insights, and mitigate against some of the limitations of experiments discussed in [Section 4](#).

### **3. SELECTED OVERVIEW OF FIELD EXPERIMENTS IN EDUCATION IN DEVELOPING COUNTRIES**

As discussed earlier, this section does not aim to provide a comprehensive review of field experiments in education in developing countries (see [Glewwe and Muralidharan, 2016](#) for such a treatment), but rather aims to illustrate the breadth of topics studied in this literature, and the broad patterns in the results to date. To organize the discussion below, I classify the range of interventions discussed into four broad categories: (1) those that are intended to increase the demand for schooling by students and their parents; (2) those that provide standard educational inputs through schools; (3) those that are related to changes in pedagogy; and finally (4) those that are related to the governance of schools, and to education systems more broadly. In the discussion below, I use the terms “experiment” and “RCT” interchangeably.

#### **3.1 Demand-side interventions**

The logic of demand-side interventions to improve education outcomes is that households may suboptimally demand too little education for their children. Reasons include not accounting for spillovers from the education of their children to the overall economy, discounting the future at a higher rate than a social planner, having incorrect beliefs about the returns to education, and being credit-constrained and unable to borrow for education even though investing in education would have a positive return. Thus,

demand-side interventions aim to correct some of these sources of suboptimal education investments.

Perhaps the most widely studied demand-side intervention using RCTs has been the idea of “conditional cash transfers (CCTs)” (with eligibility often targeted to poorer households) whereby households receive a regular cash supplement if their children are enrolled in school and maintain a minimum attendance rate. While CCT programs aim to provide income support to the poor more generally (and not just increase demand for education), they have been found to have significant positive impacts on school enrollment and attendance across most settings where they have been evaluated using an RCT (see [Fiszbein and Schady, 2009](#) for a review of several of the early studies).

RCTs have also been used to study whether modifying the design of cash transfer programs can improve (or expand) the impacts of these initiatives. For instance, [Baird et al. \(2011\)](#) study the importance of conditioning cash transfers on school enrollment by comparing a standard CCT to an unconditional cash transfers (UCTs) in Malawi and find that CCTs increase school enrollment by a larger amount than UCTs, but UCTs do better at protecting vulnerable girls by providing them with income even if they drop out of school. Similarly, [Benhassine et al. \(2013\)](#) find that labeling a UCT as being for education (in Morocco) achieved significant gains in school participation, and that adding conditionality did not yield any additional gains in schooling (though it added additional costs of monitoring and enforcing the conditionality). Finally, [Barrera-Osorio et al. \(2011\)](#) find that postponing part of the monthly transfer of a CCT to the time when school re-enrollment has to take place (which is when fees need to be paid) significantly raised enrollment relative to a standard CCT in Colombia.

In addition to studies evaluating the impact of CCTs on school participation, the randomized rollout of CCT programs across individuals and communities (most notably PROGRESA-*Oportunidades-Prospера* in Mexico) has also enabled well-identified studies on important determinants of education participation including peer effects ([Bobonis and Finan, 2009](#); [Lalive and Cattaneo, 2009](#)), consumption smoothing across households within communities ([Angelucci and De Giorgi, 2009](#)), and the role of income controlled by women/mothers on children’s consumption and education ([Bobonis, 2009](#)). Finally, [Todd and Wolpin \(2006\)](#) and [Attanasio et al. \(2012\)](#) combine a structural model with PROGRESA’s experimental variation to generate predictions on the schooling impact of the program under different values and design of the CCT program. Overall, CCTs have been one of the most highly researched and deployed policy options to improve demand for schooling in developing countries, and have been a poster child for the value of carefully randomized program rollouts in generating high-quality evidence on both program impact, as well as on deeper determinants of household schooling investments.

A second prominent category of demand-side interventions that have been studied experimentally relates to the provision of better information about education to students

and parents. Since education decisions are taken on the basis of *perceived* as opposed to actual returns (Majumdar, 1983), households may make suboptimal decisions on education investments if they misperceive these returns. Jensen (2010) uses household survey data in the Dominican Republic to show that the perceived returns to high school are much lower than the actual returns, and demonstrates experimentally that simply providing students in randomly selected schools better information on the higher measured returns to secondary schooling led to a significant increase in the years of school completed. In a variant of this experiment, Jensen (2012) shows that providing information on the job opportunities available to educated young women and helping them access these opportunities (without changing the qualifications or standards for being hired) in randomly selected villages in northern India led to a significant increase in female education, and to delays in marriage and fertility in these villages.

Finally, Loyalka et al. (2013) conduct an experimental evaluation of the impact of providing information on returns to education, and career-counseling services (in separate nonoverlapping treatments) to junior high-school students in China. They find that the information treatment had no impact on high-school participation, and also find that the career counseling treatment actually increased school dropouts and reduced test scores. The authors attribute this surprising negative result to the fact that wages of unskilled workers were rapidly rising in China in this period, and the possibility that the counseling services may have made academically weaker students decide that the academic requirements of higher education were too onerous and that it may make more sense for them to drop out and join the labor force.<sup>6</sup> The difference in the results of similar interventions across country-contexts highlights the importance of caution in comparing results across contexts. It is also important to recognize that the impact of information will likely vary as a function of the prior beliefs and the extent and direction in which the information moved these beliefs.

RCTs have also been used to study the impact of providing information to students and parents about students' learning levels, with the idea being that parental- and student investments in education is a function of their beliefs about the student's academic ability and that a misperception of true ability may lead to suboptimal education-investment decisions. Prominent examples include Dizon-Ross (2016) in Malawi and Bobba and Frisáncho (2016) in Mexico. Both studies find evidence of mismatch as well as evidence of behavioral responses to the provision of information that is consistent with households updating their decisions in response to changes in their beliefs.

RCTs have also been used to study the impact of providing information on school quality in competitive education markets with multiple providers (both public and

<sup>6</sup> Note that this result is similar to that seen nonexperimentally in Atkin (2016) who shows that Mexican high-school students were more likely to drop out from school during a period of increasing demand for unskilled labor.

private). [Andrabi et al. \(2015\)](#) use a large-scale RCT across 112 villages in Pakistan to study the impact of providing parents with detailed student and school-level report-cards with information on test scores. They find that the intervention increased mean test scores by  $0.11\sigma$ , reduced mean private school fees by 17%, and also increased primary school enrollment by 4.5%. They also find that the mechanism for these results was an improvement in quality among the lower quality schools and a reduction in price among the higher quality schools, which is consistent with the predictions of models of optimal endogenous pricing and quality choice by providers in settings of asymmetric information (and how these should change in response to provision of better market-level information on quality)<sup>7</sup>. This study highlights the capacity of RCTs to yield *market-level* insights on how the provision of information can affect parental demand and increase competitive pressure on schools and change the outcomes.

A final category of demand-side interventions with promising experimental evidence on positive impacts is the provision of student-level incentives for better academic performance. Two prominent studies include [Kremer et al. \(2009\)](#), and [Blimpo \(2014\)](#). The first study conducts an RCT of a girl's merit scholarship in Kenya and finds significant positive effects on girls' test scores and reductions in teacher absence in treatment schools. Similarly, [Blimpo \(2014\)](#) conducts an RCT of three different types of student incentives in Benin (one based on individual incentives and two based on team incentives), and finds that all three variants had a significant positive impact on the high-school exit-exam test scores. Finally, [Hirshleifer \(2015\)](#) presents experimental evidence of the relative impact of rewarding students on the basis of education inputs (measured by their performance on practice exercises) versus rewarding them on the basis of education outputs (measured by their test scores), and finds that students with input rewards significantly outperform those with output rewards.

At the same time, it is important to note that demand-side interventions are not uncontroversial. For instance, the provision of information may make recipients of the information worse-off if it is not correct. Specifically, one concern with the approach in [Jensen \(2010\)](#) is that the information provided on "Mincerian returns to education" may be incorrect on average (due to omitted variable bias) and also not be correct for the marginal student, since the returns to education for the marginal student induced to stay in school by the intervention are likely different from those to the average student ([Carneiro et al., 2011](#)).<sup>8</sup> Similarly, opponents of student incentives express concern that rewarding students for test-score gains may crowd out students' intrinsic motivation for learning and acquiring knowledge for its own sake.

<sup>7</sup> They also verify that parents' knowledge of school quality did improve as a result of the intervention.

<sup>8</sup> Note that the approach in [Jensen \(2012\)](#) is less susceptible to this concern because the standards for hiring candidates did not change (and hence no potentially misleading information was provided) but new information was provided by the recruiting firms.

More generally, demand-side interventions tend to be paternalistic in nature since they typically assume that households are making suboptimal education choices and need to be induced to demand more education. On the one hand, there is considerable evidence that there are important demand-side market failures that may lead to suboptimal investments in education by parents and children. For instance, the large positive impacts of relatively small student prizes and incentives (which are especially small relative to the lifetime returns to completing schooling) suggest that students may underestimate the returns to education (or discount the future at a significantly higher rate than a social planner would). On the other hand, it is also possible that seemingly suboptimal choices may be optimal in a local context and that a well-intentioned demand-side intervention may make people worse-off. Thus, it is a good practice for designers of demand-side interventions to check that they are responding to demonstrated evidence of suboptimal choices in the specific context where the intervention is being considered, as opposed to simply assuming that this is true.

Overall, the evidence from experimental evaluations on the impact of demand-side interventions suggests that well-designed demand-side interventions are likely to be a promising avenue for improving education outcomes in developing countries. Some interventions like the provision of better information are inexpensive and easy to scale up. Others like CCTs are expensive and much less cost-effective in terms of the additional years of schooling obtained per dollar spent ([Dhaliwal et al., 2012](#)).<sup>9</sup> Key open questions include experimenting with alternative designs of demand-side interventions to better match the intervention to the source of inefficiency and doing so in the most cost-effective manner.

### **3.2 School and student inputs**

The vast majority of public education spending is devoted to school inputs including infrastructure, teacher salaries, and student inputs including textbooks and other learning materials. A considerable amount of education research has aimed to study the impacts of these inputs on school participation (enrollment and attendance) as well as learning outcomes (see [Glewwe and Muralidharan, 2016](#) for a review of this research).

However, there are relatively few RCTs of school construction and infrastructure construction since it is not easy to randomize construction of durable assets. [Bурde and Linden \(2013\)](#) experimentally vary the creation of village-based schools in rural Afghanistan and find large effects of having a school in the village on school participation and test scores, especially for girls. They also use the exogenous variation in distance to

<sup>9</sup> This is mainly because the transfers are also provided to inframarginal households who would have sent their children to school regardless of the CCT. On the other hand, note also that CCTs are broad social protection programs that aim to achieve several goals other than improving education and are perhaps by design not as cost-effective at improving the education outcomes.

the nearest school induced by their experiments and estimate that the distance-elasticity of school attendance is quite large, again, especially for girls.<sup>10</sup> Thus, school construction, particularly in villages that previously had no schools,—is likely to improve enrollment (and potentially test scores as well), relative to the counterfactual of not attending school. On the other hand, it is also true that the construction of schools in every village (or even hamlets within large villages) has led to the creation of many subscale schools where low enrollments lead to teachers typically teaching multiple grades at the same time (especially in India). Thus, a key open question with regard to school access is understanding the optimal trade-off between access and scale; in particular, whether it is better to have fewer, larger schools that are better equipped and managed, with transport subsidies to ensure access to students living outside a walking distance to the school.<sup>11</sup>

Unlike infrastructure, there have been several experimental evaluations of the impact of providing schools with books and materials (or grants to be used for books and materials), and a consistent finding across studies has been that the provision of these materials *on their own* does not lead to significant improvements in either school participation or learning outcomes. Six of these studies are briefly summarized below.

Glewwe et al. (2009) conduct an RCT evaluating the provision of free textbooks to school children in Kenya and find that this had no impact on either student attendance or learning outcomes. Sabarwal et al. (2014) experimentally evaluate a similar program in Sierra Leone and again find no impact on attendance or learning outcomes. Das et al. (2013) present results from an experimental evaluation of the provision of a block grant to schools in the Indian state of Andhra Pradesh (that was mainly used to buy books and materials) and find no impacts on test scores after two years of the program. Mbiti et al. (2016) conduct an experimental evaluation of a school grant program in Tanzania, which was again mainly used for textbooks and learning materials and also find no impacts on test scores after two years of such grants. Borkum et al., 2013 experimentally evaluate a program, in the Indian state of Karnataka, that provided schools with libraries (consisting of providing a collection of books and a librarian rather than physical construction of a library, which makes the intervention closer to the one that provided books rather than infrastructure) and find no impact of the program on learning outcomes. Finally, Pradhan et al. (2014) evaluate the impact of a community engagement program to improve schooling in Indonesia where they include a treatment arm that provided a block grant as a benchmark against which to study the other interventions, and find that the provision of the block grant had no impact on learning outcomes. While the

<sup>10</sup> These experimental findings are consistent with those from well-identified studies using difference-in-difference methods to study the impact of new school construction on enrollment (Duflo, 2001) or on providing subsidized transport to improve access to school (Muralidharan and Prakash, 2016).

<sup>11</sup> For instance, the Indian state of Rajasthan recently took a policy decision toward such a consolidation, though there has been no evaluation to date of its impacts.

reasons for the zero effect may vary across specific studies and contexts (see discussion in [Section 4.2](#)), the breadth of the evidence above, spanning Africa, South Asia, and South-east Asia, suggests quite strongly that simply providing more resources to schools is unlikely to improve learning outcomes.

The broad theme that simply providing inputs to schools and students may have limited impacts on learning comes is corroborated by evidence on a different class of educational inputs, namely computers. A striking example of this is provided by [Cristia et al. \(2012\)](#) and [Beuermann et al. \(2015\)](#) who conduct a large experimental evaluation of the “One-Laptop-per-Child” program in Peru, and find that even though the program led to a sharp increase in the fraction of children with access to a computer, it had no impact on learning outcomes. In contrast, interventions that effectively utilize technology to improve pedagogy typically have positive impacts on test scores (see [Section 3.3](#)).

Teachers are another critical input into education and teacher-salaries account for the majority of education-spending in most countries. The key research questions of interest are studying the impact of teacher quantity and teacher quality on education outcomes. In practice, the number of teachers hired is determined by the average class-size norms that school systems try to achieve, and so the key research question of interest is the impact of class size (also referred to as pupil–teacher ratio) on learning outcomes.

The experimental evidence on class size in developing countries is limited because it is not easy to randomly assign civil-service teachers across schools. However, there is indirect experimental evidence on the impact of class size from multiple studies. The best evidence comes from the “Extra Teacher Project” analyzed in [Duflo et al. \(2011, 2015b\)](#). The project conducted an RCT in Kenya that randomly assigned some schools an extra contract teacher. This extra teacher was assigned to grade 1 and used to reduce class size by roughly 50% (halving the pupil-teacher ratio from roughly 80:1 to 40:1). Further, half the students in first grade in treatment schools were randomly assigned to classes taught by current (civil-service) teachers, and the other half were assigned to classes taught by contract teachers. For the purpose of identifying the impact of the pupil-teacher ratio on student learning the classes taught by the current (civil-service) teacher can be compared to those taught by same type of teacher in the control schools, which have much larger pupil-teacher ratios. [Duflo et al., 2015b](#) report that although this reduction in class size led to higher test scores (about  $0.09\sigma$ ), this increase was not statistically significant.

Another piece of indirect evidence is provided by [Banerjee et al. \(2007\)](#) who conduct an experimental evaluation of a remedial education program where students with low test scores were “pulled out” of class for remedial instruction for a few hours each day. While the test scores of the students who received this remedial instruction went up significantly, the authors find no impact on test scores of students who remained in the original classes with smaller class sizes, suggesting that smaller class sizes may have limited impact on improving learning outcomes.

On teacher quality, there is limited experimental evidence on the impact of teacher-training programs although panel-data-based studies typically find no correlation between possession of a teacher-training credential and measures of teacher value-added ([Muralidharan, 2013](#)). The experimental study that most closely resembles a study of teacher training is in [Muralidharan and Sundararaman \(2010\)](#) who study the impact of providing teachers with detailed written diagnostic feedback on the performance of their students on external assessments along with suggestions for more effective teaching, and find that there was no impact whatsoever of the program on learning outcomes (not only were the point estimates of impact zero, but the distributions of test scores in treatment and control groups were nearly identical). In interpreting these results, the authors present complementary evidence to suggest that the zero impact was not because the content of the diagnostic feedback was not useful, but rather that teachers in government-run public schools did not have the motivation or professional incentives to use this feedback and revise their teaching practices.<sup>12</sup>

Finally, teacher salaries are often considered an important source of teacher motivation, and many global education advocates recommend raising teacher salaries to improve their motivation and effort. Reflecting this thinking, the Government of Indonesia passed an ambitious new teacher reform in 2005 that permanently doubled the base pay of teachers who passed a simple certification process. Using a large-scale experiment that accelerated the doubling of pay for teachers in randomly selected schools, [de Ree et al. \(2015\)](#) show that the program significantly improved teacher satisfaction with their income, reduced the incidence of teachers holding outside jobs, and reduced self-reported financial stress. However, despite these changes in measures of teacher's well-being and satisfaction, there was no impact on either teacher effort or student learning (the study focused on the impacts on incumbent teachers and not on the long-term impact on quality of new teachers).

In contrast with the mostly negative results summarized so far, one class of input-based interventions that have shown promising results are those that improve student health. The most striking example of this is the positive impact (both short- and long-term) of school-based deworming treatments in settings with high rates of child worm infections. [Miguel and Kremer \(2004\)](#) provide experimental evidence on the impact of a school-based deworming program in Kenya and find that it led to a significant improvement in school attendance in treated schools. While they did not find impacts on test scores in the short run, a follow-up study found significant long-term positive impacts from exposure to the program. [Baird et al. \(2016\)](#) report results from the long-term

<sup>12</sup> Specifically, they show that the extent to which teachers report that the feedback reports had useful content was not correlated with student test-score gains in the schools that only received the feedback reports, but that it was significantly positively correlated with student test-score gains in schools that received both feedback reports and performance-linked bonuses for teachers. Thus, when teachers were rewarded for improving student test scores, it appears that they did make use of the diagnostic feedback reports, but not otherwise.

follow-up and find that 10 years after the treatment, men who were eligible for the treatment when they were boys were more likely to be enrolled in school, worked 17% longer hours each week, and missed one fewer meal per week. Women who were eligible for the program when they were girls were more likely to have attended secondary school. The authors estimate that the annual internal rate of return of the deworming program was 32%, making it a highly cost-effective program.

Overall, the experimental evidence on various categories of school inputs suggests that most of the “business-as-usual” patterns of expenditure seem to be rather ineffective at improving education outcomes. Why might this be the case? As I discuss further in [Sections 3.3 and 3.4](#), there is considerable evidence to suggest that providing traditional inputs to schools does not typically alleviate binding constraints (of pedagogy and governance) to improving learning outcomes in many low-income settings. Thus, better inputs *may* have a positive impact in settings where these other binding constraints have been alleviated, but appear to have more limited impact on learning outcomes in developing country settings where constraints in pedagogy and governance are more first order.

### 3.3 Pedagogy

While education inputs (infrastructure, teachers, and materials) typically account for the majority of education expenditure, a critical determinant of the extent to which these inputs translate into better learning outcomes is the pedagogy in the classroom. In other words, the “technology of instruction” is a key determinant of how inputs are translated into outcomes, and improvements in this “technology” may yield substantial improvements in outcomes. Researchers have only recently started to use experimental techniques for identifying the impacts of pedagogical innovations on learning outcomes, and there is a relative paucity of experimental evidence on the critical question of how best to structure classroom instruction and pedagogy to most effectively improve learning outcomes in developing countries. Nevertheless, considerable progress has been made on obtaining a better understanding of these issues in the past decade, which in turn has provided important insights into the nature of the education-production function, and the binding constraints that confound attempts to improve education outcomes in developing countries.

The most important insight on pedagogy that has been obtained in the past 15 years of experimental research on education in developing countries is that their education systems are poorly equipped to handle the pedagogical challenges posed by tens of millions of first-generation learners entering formal schooling. In particular, the curricula and teaching practices that may have been optimal at a time when education was more limited have not been adapted to deal with the challenge posed by increasing heterogeneity in classroom preparedness. Also, as [Muralidharan and Zieleniak \(2014\)](#) show, the variance in student-learning levels within a cohort increases over time as they pass through the school system.

Thus, designing effective pedagogical strategies for handling variation in academic preparation across students in a classroom is a fundamental challenge for effective teaching and learning, and the experimental evidence over the past 15 years has consistently found large positive effects of interventions that aim to address this challenge. We highlight three categories of interventions: supplementary teaching at the right level (TaRL), classroom tracking, and individually customized computer-aided learning (CAL).

The pedagogical intervention with the most consistent evidence of strong positive effects is that of supplemental education that ignores the text book (of the grade the student is enrolled in) and instead focuses on teaching children “at the right level”. In practice, this means that children who are unable to read are taught the alphabet; those who can read alphabets are taught to put them together to read words; those who can read words are taught to read sentences, and so on. These programs (many of them designed by *Pratham*, a nonprofit organization focused on education based in India) are typically delivered by young women who have completed secondary school or high school who do not have any formal teacher training credentials and who are paid very modest stipends.

Several RCTs of *Pratham*'s TaRL program have been carried out over the past decade across different locations and implementation models and the effects of these interventions have been positive in many settings ranging from urban locations in Western India (Banerjee et al., 2007) to rural North India (Banerjee et al., 2010, 2016; Duflo et al., 2015a). Further evidence in favor of this approach is provided by Lakshminarayana et al. (2013), who study the impact of a program run by a different nonprofit (the Naandi Foundation) that recruited community volunteers to provide remedial education to children in a randomly selected set of villages in the Indian state of Andhra Pradesh, and find that student test scores in program villages were  $0.74\sigma$  higher than those in the comparison group after two years.

More recently, Duflo et al. (2015a), and Banerjee et al. (2016) present results from multiple RCTs conducted across several Indian states (Bihar, Uttarakhand, Haryana, and Uttar Pradesh) in partnership with *Pratham* to evaluate the impact of different models of implementing the TaRL approach in public schools. Overall, they find support for the hypothesis that *Pratham*'s instructional approach, which focuses on teaching children at a level that matches their level of learning, can significantly improve learning outcomes. However, they find considerable variation in the effectiveness of different implementation models. They find that implementing the pedagogy in summer camps that are held outside of normal school hours or in dedicated learning camps in the school (at which point children were grouped by their level of academic preparation as opposed to by grade level) was highly effective in raising test scores.<sup>13</sup> However,

<sup>13</sup> Note, however, that the model within-school learning camps was much more effective overall because student attendance at the summer camps was limited, whereas student-attendance rates were much higher at the in-school learning camps (see the further discussion in the conclusion regarding scaling up successful interventions).

they found that it was more difficult to achieve large gains under implementation models that attempted to incorporate this pedagogy into the teaching by regular teachers during the school day.

For example, the first attempt to scale up the TaRL model in public schools in Bihar and in Uttarakhand consisted of training regular teachers in the TaRL pedagogy and providing instructional materials to support implementation, but these programs did not lead to any improvement in learning outcomes. In Bihar, a variant of the program that also had volunteers providing supplemental instruction outside the school using the TaRL method improved language and math scores by  $0.11\sigma$  and  $0.13\sigma$ . However, in Uttarakhand, even providing an extra volunteer did not improve test scores because the volunteers were provided within schools and were used by teachers to implement regular as opposed to TaRL pedagogy. Recognizing this challenge, the design of the TaRL program in Haryana relied on schools dedicating an hour a day for the TaRL pedagogy, where students in primary schools were reorganized on the basis of learning levels as opposed to the grade they were enrolled in. [Duflo et al. \(2015a\)](#) evaluate this implementation model and find that it improved reading scores by  $0.15\sigma$ . Finally, the most successful implementation model was that of “learning camps” conducted within government schools in Uttar Pradesh, where Pratham volunteers essentially took over the running of schools for 50 days (through five 10-day camps including one in the summer) and achieved remarkably large gains of  $0.7\sigma$  in both math and language.

The authors interpret their findings as suggesting that the remedial pedagogy was successful, but that it was difficult to get teachers to implement a new curriculum during school hours, and that successfully scaling up remedial pedagogy within an existing schooling system can be challenging because teachers are focused on completing the syllabus prescribed in the text book. Successful models required either dedicated reorganizing of school time committed to by the government (as in Haryana) or the taking over of school time by volunteers focused on implementing the TaRL model (as in Uttar Pradesh).

A second way of reducing the variance in student preparedness within a classroom is to “track” students into streams within a grade as a function of their preparedness. While proponents argue that tracking would benefit all students by allowing teachers to better tailor their levels of instruction, opponents are usually concerned that students who are tracked to “lower” level classrooms may suffer further from negative peer effects, stereotyping, and loss of self-esteem, which may place them on a permanently lower trajectory of learning.

[Duflo et al. \(2011\)](#) provide important evidence on this question by experimentally evaluating a program in Kenya that tracked students in first grade into two classrooms based on their baseline test scores. They found that students in tracked schools had significantly higher test scores ( $0.18\sigma$ ) than students in nontracked schools, and that they

continued to score  $0.18\sigma$  higher even one year after the tracking program ended, suggesting longer-lasting impacts than those found in many other education interventions. They also found positive impacts for students at all quartiles of the initial test score distribution and could not reject that students who started out above the median score gained the same as those below the median; Additionally, lower-achieving students gained knowledge in basic skills, while higher-achieving students gained knowledge in more advanced skills, suggesting that teachers tailored their classes to the achievement level of their students. Finally, since students just below- and just above the median baseline score were otherwise similar, but experienced a sharp change in the mean test score of their peers, the authors are able to use this regression discontinuity method to show that tracking did not cause adverse peer effects in this setting.

A third way of differentiating instruction for students at different levels of preparation is computer-adaptive learning. While there are several reasons to be optimistic about the *potential* for technology to improve learning outcomes (including the ability to overcome limitations in teacher knowledge, tailor instruction to students' needs, and provide real-time feedback to students), the evidence, to date, on the impact of interventions that simply provide computer hardware suggests zero-to negative impacts on test scores (Barrera-Osorio and Linden, 2009; Cristia et al., 2012; Beuermann et al., 2015; Malamud and Pop-Eleches, 2011). On the other hand, interventions that focus on using technology for better pedagogy have typically found more positive results.

Banerjee et al. (2007) evaluate a CAL program that consisted of math games that were designed to emphasize basic competencies in the official math curriculum in two cities in Western India, and find large gains in test scores at the end of one- and two years of the program ( $0.35$  and  $0.47\sigma$ , respectively). Further, a series of experimental evaluations of CAL in China have found modest positive impacts on learning (in the range of  $0.12$  to  $0.25\sigma$ ). These studies include Lai et al. (2015) who study a CAL program in schools for migrant children in Beijing; Yang et al. (2013) who study a CAL program in three provinces in China (Shaanxi, Qinghai, and Beijing) in schools for socioeconomically disadvantaged students. However, the CAL programs in the experiments in China used technology to reinforce grade-appropriate content and did not feature extensive individual customization, and the CAL program in India featured children sharing computer time, which may have limited the ability of the system to provide individual customization.

Recent evidence on the potentially large benefits of individually customized CAL programs is provided by Muralidharan et al., 2017b who experimentally evaluate a CAL program (called *Mindspark*) that was explicitly designed to customize pedagogy to the right level of students in grades 6 to 9 in New Delhi, India. The *Mindspark* program was developed iteratively over many years by a leading education diagnostic testing firm in India and featured a question bank of over 45,000 individual questions calibrated finely

to students' existing achievement levels using over two million observations of student-item responses. Further, the *Mindspark* system analyzes patterns of student responses to the screening test and algorithmically identifies areas of conceptual misunderstanding. Thus, the computer-adaptive learning system was able to finely calibrate students' baseline competencies and tailor the academic content to those levels. Further, the system dynamically adjusts the content that students are working on at any given point of time based on their response to the previous question. This ability to customize instruction may be especially important in post-primary grades where the variation of student ability is likely to be even larger than in primary grades.

Using detailed question-level data, [Muralidharan et al. \(2017b\)](#) report five results. First, students in this setting are several grade levels behind their grade-appropriate standard, and this gap grows by grade. Second, in the control group, students in the lower half of the baseline test-score distribution in a grade show *no improvements* in test scores during the school year, suggesting that the level of the “business as usual” curriculum may be too high for students in the lower end of the baseline test-score distribution to learn anything effectively. Third, they estimate that attending Mindspark centers for 90 days would increase student test scores by  $0.59\sigma$  and  $0.36\sigma$  in math and Hindi (language), respectively. Fourth, consistent with the promise of customized instruction, the treatment effects are equally large at all levels of absolute baseline scores. Fifth, because the “business-as-usual” rate of learning for academically weaker students is close to zero, their relative improvement from the technology-aided curriculum is much greater.

These results suggest that curricular mismatch may limit the extent to which time in school translates into additional learning in developing countries (especially for students with low foundational skills), and that CAL programs designed to adjust to the level of the student can sharply improve learning outcomes for all students in such settings.

The main insight from the evidence summarized in this section is that a key binding constraint in converting inputs into learning outcomes in developing countries may be the fact that “business-as-usual” instructional practices that follow the textbook and curriculum are not at the correct level for the majority of students in developing country education systems. Attempts to address this problem that have been successful to date include supplemental remedial instruction, tracking, and customized computer-adaptive learning tools.

Thus, investing in designing effective pedagogy to handle the large variation in student preparation in developing country classrooms is likely to yield considerable positive returns in these settings. Further, in addition to designing and evaluating technical solutions for better pedagogy, considerable additional efforts are required to embed these improvements in systems to improve pedagogy at scale. As the evidence

in [Banerjee et al. \(2016\)](#) shows, it is not easy to change default teaching practices to incorporate evidence on new and more effective pedagogy. These are all areas where more experimentation and research are needed.

### 3.4 Governance

A fourth critical determinant of education outcomes in addition to household demand, the provision of inputs, and the details of classroom pedagogy is the overall governance of the education system. I use the term “governance” broadly to include goal-setting for the education system, motivating service providers to deliver on these goals and holding them accountable if they do not, and the quality of management of schools and school systems.

One striking indicator of weak governance in schools in developing countries is the high rate of teacher absence. [Chaudhury et al. \(2006\)](#) present results from a multicountry study where enumerators made unannounced visits to public schools to measure teacher attendance and activity, and report an average teacher absence rate of 19%, with teacher absence rates of 25% in India and 27% in Uganda. [Muralidharan et al. \(2017a\)](#) present more recent estimates from a nationally representative panel survey that revisited villages surveyed by [Chaudhury et al. \(2006\)](#), and find only a modest reduction in teacher absence rates—from 26.3% to 23.7%. They also calculate that the fiscal cost of teacher absence is \$1.5 billion *each year*, highlighting the large costs of poor governance in education. Another example of weak governance is direct corruption in education spending where large amounts of funds meant for schools simply do not reach them. For instance, [Reinikka and Svensson \(2004\)](#) show that 87% of central government funds allocated to a school capitation grant (for nonwage expenditures) in Uganda never reached the schools, and that the median school in their representative sample had not received *any* of the funds.

A considerable body of experimental evidence has been accumulated over the past decade on the impact of interventions to improve governance and management of education systems in developing countries. The broad summary of this literature is that while there is variation in the effectiveness of individual interventions, there is enough evidence to suggest that there are highly cost-effective interventions that can substantially improve governance, if there is political willingness to do so. In other words, the evidence suggests that *technical* solutions exist for improving governance, but that scaling these up may be difficult without the political desire to do so. I briefly summarize four categories of interventions to improve school governance: decentralizing more authority to communities for school management, modifying teacher contractual structure, performance-linked pay for teachers, and private management/vouchers.

The theory of change behind decentralizing school management is that providing more authority to communities to hold schools and teachers accountable for performance will improve teacher effort and student learning outcomes. However, the evidence from five different RCTs of such policy changes in different settings suggest that this approach may not be very effective. [Banerjee et al. \(2010\)](#) present experimental evidence on a program in rural North India that tried to empower communities to oversee their schools by making them more aware of their rights over schools and found no impact on learning outcomes. [Pradhan et al. \(2014\)](#) find limited effects of enhancing community participation in school governance in Indonesia. [Beasley and Huillery \(2012\)](#) experimentally evaluate a program providing grants to school committees to encourage parental participation in school management in Niger and find no impacts on test scores. Finally, [Lassibille et al. \(2010\)](#) and [Glewwe and Maïga \(2011\)](#) both present experimental evaluations of the AGEMAD program in Madagascar that aimed to strengthen school management at the district, subdistrict, school, and teacher levels, and find no impact on student test scores of these interventions.

A likely explanation for these results is that communities in practice do not have much authority over teachers, who are typically civil-service employees and also typically much more educated (and hence powerful) than the residents of the communities that they serve. For instance, [Pradhan et al. \(2014\)](#) find that though having elections for school committees did not improve outcomes on its own, it did improve learning outcomes when implemented as part of a treatment that also provided formal linkages to the village council through joint planning meetings. The authors argue that this was because the village council was more powerful and that the linkage provided greater legitimacy to cosponsored activities of the village council and the school committee. In contrast, enhanced community participation alone did not provide the school committee with enough power to impact learning.

Further evidence in support of the idea that a key constraint to the effectiveness of school management committees is their authority over schools is provided by [Duflo et al., 2015b](#). They find that training the school management committees to evaluate the performance of contract teachers and to have inputs into the renewals of contract-teacher contracts had a significant positive impact on the performance of the contract teachers and on student test scores. Thus, in cases where the committees had direct authority over the renewal of teacher employment contracts (and where the teachers belonged to the same community), they were able to make a difference. Thus, improving accountability of teachers is likely to matter, but not all policy attempts (like increasing community participation) may be effective at doing so.

A second way of improving teacher accountability and performance is to modify the structure of employment contracts so that the renewal of employment is contingent on measures of performance. Of course, most teachers in both developed and developing

countries are employed in the public sector, and public sector employment contracts typically feature lifetime tenure after a short probationary period (if any). However, in recent years, many developing countries have started to employ new teachers on short-term renewable contracts. Contract teachers comprise a third of public-school teachers across 12 countries in Africa (Bourdon et al., 2010) and their share among all public-school teachers in India grew from 6% in 2003 to 30% in 2010 (Muralidharan et al., 2017a).

There are two notable experimental studies on the impact of contract teachers. First, Duflo et al., 2015b conduct an RCT of the impact of providing randomly selected schools in Kenya with an extra contract teacher, who was assigned to grade 1 and used to reduce class size in first grade by half (with students randomly assigned to either the section with a contract teacher or the one with a regular teacher). They find that class-size reductions with regular teachers did not have a significant impact on test scores, but that students with reduced class sizes who also had a contract teacher scored  $0.29\sigma$  higher than those in control schools. They also find that holding the class size constant, students taught by contract teachers scored significantly higher than those taught by civil-service teachers.

Second, Muralidharan and Sundararaman (2013) conduct an RCT of contract teachers of the impact of providing randomly selected schools in the Indian state of Andhra Pradesh with an extra contract teacher, where the schools were free to assign the teacher as they wished (since the schools featured multigrade teaching, the optimal use of the teacher would vary considerably across schools and it would be difficult to ensure fidelity to a within-school randomization design). They find that students in schools with the extra contract teacher had significantly higher test scores after two years of the program, and show using several nonexperimental techniques that the contract teachers were at least as effective as regular teachers at improving test scores. Both studies find that contract teachers had significantly lower absence rates than regular teachers, and also find that the absence rates of regular teachers increased in schools with an extra contract teacher, suggesting that the estimated effects may be a lower bound on the true effect of an extra contract teacher.

Contract teachers tend to be different from civil-service teachers in many ways (they are typically less educated, less likely to have formal teacher training credentials, more likely to be from the local community, and typically paid much lower salaries) and neither of the studies above (or any other study) can isolate the impact of just changing the contractual structure of employment holding all other factors constant. However, since most of the other differences would suggest that contract teachers may be less effective than regular teachers (such as being less educated, less trained, and paid much lower salaries), the evidence from the two experiments suggests that the renewable nature of the contracts may have contributed to the greater accountability of the contract teachers.

A third way of improving teacher effort and motivation that has been well studied is the idea of introducing performance-linked pay for teachers.<sup>14</sup> There are four noteworthy experimental studies on this topic.

First, [Muralidharan and Sundararaman \(2011\)](#) conduct an RCT of a program in the Indian state of Andhra Pradesh that paid teachers bonuses on the basis of the average improvement in test scores of their students. They find significant improvements in test scores of students in treated schools after 2 years (of  $0.27\sigma$  and  $0.17\sigma$  in math and language) and find no evidence of any negative effects. Students in treated schools did better on both “mechanical” and “conceptual” components of the test, where the former were designed to reflect questions in the text book (that could be “taught to”) while the latter were designed to reflect deeper understanding of the materials. Students in treated schools also did better on science and social studies tests (for which there were no incentives) suggesting positive spillovers from improvements in math and science.

Second, [Muralidharan \(2012\)](#) presents evidence from a 5-year long follow-up of the original experiment (where a randomly selected subset of the original schools saw the continuation of the performance-pay program for 5 years) and reports strong positive effects for the cohort whose teachers were eligible for performance pay for 5 years (their test scores were  $0.54\sigma$  and  $0.35\sigma$  higher in math and language). The study also finds that though group and individual teacher incentives appeared equally effective in the short run, the individual incentives significantly outperformed the group incentives at the end of 5 years.

Third, [Glewwe et al. \(2010\)](#) conducted an experimental evaluation of a teacher incentive program in Kenya that provided school-level group incentives using prizes for high-achieving schools, and find that students in treatment schools did score better on high-stakes tests but not on low-stakes tests, and also that these gains dissipated after the incentive program was over. They interpret their results as suggesting that teacher incentives may not be effective as a strategy for promoting long-term learning. Nevertheless, there are two important caveats. The first is that we now know that *all* interventions appear to have significantly high rates of test-score decay (see [Andrabi et al., 2011](#)), and that there may be important long-term gains in human capital even when test-score gains decay ([Chetty et al., 2011](#)). Second, the group nature of the incentive program (across 12 teachers) may have induced free riding and weakened the incentives faced by individual teachers (as seen in [Muralidharan, 2012](#)).

<sup>14</sup> There are several reasons why default compensation systems for teachers have little or no link to performance. These include difficulties in measuring productivity of individual teachers, as well as concerns that linking pay to performance on measurable attributes of a job will lead to diversion of effort away from socially valuable tasks that may not be as well measured ([Holmstrom and Milgrom, 1991](#); [Baker, 1992](#)). Nevertheless, the demonstrated low levels of teacher effort in developing countries (manifested by high rates of absence) have led both policy makers and researchers to consider the possibility that introducing performance-linked pay for teachers may improve outcomes.

Fourth, [Duflo et al. \(2012\)](#) present evidence from a program in the Indian state of Rajasthan, that paid teachers based on the number of days they attend work as opposed to a flat salary and find that the program led to a halving of teacher absence rates (from 42% to 21%) and significant increases in student test scores (by  $0.17\sigma$ ). The intervention combined both better monitoring (teacher attendance was verified with photos with time-date stamps) and better incentives (since were paid based on days of attendance). Using a structural model identified by nonlinear sections of the pay-off schedule, the authors show that the improvement in attendance is mainly attributable to the improvements in incentives as opposed to just increased monitoring by itself.

All the programs studied above featured relatively small bonuses that averaged less than 10% of monthly pay. The large positive effects from even modest amounts of pay linked to performance are particularly striking when compared with the finding of *zero* impact on student learning from an unconditional *doubling* of teacher pay in Indonesia ([de Ree et al., 2015](#)). Taken together, these results suggest that even modest changes to compensation structure to reward teachers on the basis of objective measures of performance (such as attendance or increases in student test scores) can generate substantial improvements in learning outcomes at a fraction of the cost of a “business-as-usual” expansion in education spending. However, not all performance-pay programs are likely to be effective, and so it is important to design the bonus formulae well and to make sure that these designs reflect insights from economic theory (see the discussion in [Section 5.1](#)).

The final class of governance interventions that has attracted considerable policy and research attention is the idea of combining public funding for education with competition across public and private producers of education, through voucher-based models where parents get to choose schools (public, private, or nonprofit) and the government directly reimburses the school. The promise of such voucher- and choice-based reforms is that private management may be more effective than traditional public school management and that giving parents more choice across schools (as opposed to limiting them to publicly provided schooling options) would increase competition and accountability across all schools.

School choice is a controversial subject, and experiments are particularly important in testing the relative effectiveness of public and private schools, because cross-sectional differences in test scores are likely to be confounded by omitted variables. There are two sets of well-identified studies of school voucher programs in developing countries that defrayed the cost of attending private schools. [Angrist et al. \(2002, 2006\)](#) study the short- and medium-term effects the PACES program in Colombia that provided vouchers (allocated by lottery) to students to attend private schools, and find that voucher winners scored significantly better both 3 and 7 years after receiving the voucher. However, the PACES program also allowed vouchers to be topped up by parents (to attend a better school than they could have afforded without a voucher), and required students to maintain minimum academic standards to continue receiving the voucher. Thus, while these

results point to the effectiveness of the PACES program, the estimates reflect a combination of private school productivity, additional education spending, and student incentives.

[Muralidharan and Sundararaman \(2015\)](#) present experimental evidence on the impact of a school-choice program in the Indian state of Andhra Pradesh that featured a unique two-stage randomization of the offer of a voucher (across villages as well as students). The design created a set of control *villages* that allowed the authors to experimentally evaluate both the individual impacts of school choice (using the student-level lottery) as well as its aggregate effects including the spillovers on nonapplicants and students who start out in private schools (using the village-level lottery). At the end of 2 and 4 years of the school choice program, they find no difference between the test scores of lottery winners and losers on the two main subjects of Telugu (the native language of Andhra Pradesh) and math, suggesting that the large cross-sectional test-score differences in these subjects across public and private schools (of  $0.65\sigma$ ) mostly reflect omitted variables.

However, they find that private schools spend significantly less instructional time on Telugu (40% less) and math (32% less) than public schools, and instead spend more time on English, and science and social studies. Private schools also taught a third language, Hindi, which was not taught in public primary schools. When they conduct tests in these additional subjects after 4 years of the voucher program, they find small positive effects of winning the voucher on English ( $0.12\sigma$ ;  $p$ -value = .098), and science and social studies ( $0.08\sigma$ ;  $p$ -value = .16), and large, positive effects on Hindi ( $0.55\sigma$ ;  $p$ -value < .001). Further, the annual cost per student in the public-school system is over three times the mean cost per student in the private schools in the sample.

Thus, on the one hand, private schools were clearly *more productive* than public schools (they achieved similar results on the main subjects at much lower cost, and produced gains on an additional subject that was not taught in the public schools), but they were also *not more effective* at improving learning outcomes on the core subjects. The results suggest that private management may have the potential to deliver better learning outcomes at comparable costs, but there is no evidence yet that this is the case. Thus, a key open question for future research is to study the relative effectiveness of private and public management holding the spending per student constant.<sup>15</sup>

### 3.5 Summary of evidence

As mentioned earlier, the review above does not aim to be a comprehensive review of all field experiments in education in developing countries (see [Glewwe and Muralidharan, 2016](#)

<sup>15</sup> The recently announced initiative by the Government of Liberia to launch the “Partnership Schools for Liberia” initiative provides a promising opportunity to answer this question.

for a recent summary). Rather, it aims to synthesize the consistent patterns in the evidence and highlight the most important general insights obtained from field experiments in education in developing countries in the past decade.

The key messages from this evidence are that “business-as-usual” expansion of spending on school inputs (which is where the majority of education spending is allocated) may have only modest impacts on improving education outcomes. The main reason for this appears to be that the binding constraints to better performance of developing country education systems may not be inputs but rather: (1) outdated pedagogy that focuses on completing textbooks without accounting for the fact that millions of new first-generation learners may be considerably behind the levels assumed by the textbook; a problem that gets worse in older grades, and (2) weak governance with poor accountability for teachers and other front-line service providers. Thus, these appear to be the most important areas to focus attention on, in addition to designing effective demand-side interventions.

Nevertheless, as the discussion of evidence shows, not all interventions to improve demand, pedagogy, or governance are equally effective, which underscores the need for ongoing high-quality evaluations of these initiatives. The next sections aim to provide guidelines for researchers on how to effectively design and implement such evaluations in education in developing countries.

## 4. LIMITATIONS OF FIELD EXPERIMENTS AND STRATEGIES FOR MITIGATING THEM

In this section, I provide a discussion of the important limitations of field experiments. Many of these limitations apply to almost *all empirical research that tries to identify causal relationships*, and should not be seen as weaknesses of experimental methods in particular. But it is important to be clear about what problems experiments do and do not solve, and doing so can improve the quality of policy-relevant inference made from individual studies, as well as help guide future research in ways that mitigate these challenges. The goal of discussing these limitations here is to (1) provide the necessary nuance and caveats in interpreting the results discussed in [Section 3](#), and (2) to motivate the discussion in [Section 5](#) where I describe ways of addressing these limitations through better design and data collection.

### 4.1 Production function versus policy parameters

The discussion in [Section 2.3](#) highlighted the value of experimentally varying  $X_{i,t}$  in estimating the causal impact of  $X_{i,t}$  on education outcomes. Note, however, that even random assignment of  $X_{i,t}$  may not yield the production function parameter  $\beta$  outlined in Eq. (1). This is because the production function parameter  $\beta$  is a partial derivative ( $\partial T_{i,t} / \partial X_{i,t}$ ) holding *other inputs constant*. In practice, other inputs at the school or

household level may endogenously respond to exogenous changes in  $X_{i,t}$ , and the estimated parameter should therefore be more accurately interpreted as a policy parameter, which is a total derivative ( $dT_{i,t}/dX_{i,t}$ ) that accounts for reoptimization by agents in response to an exogenous change in  $X_{i,t}$ .

The extent to which an experimental estimate reflects reoptimization will depend critically on the duration of the study. A clear illustration is provided by Das et al. (2013), who study a randomly assigned school grant program in India over a two-year period and find significant positive effects on test scores at the end of the first year, but find no effect in the second year even though the grant was provided again in the second year, and was spent on very similar items in both years (books, school supplies, and classroom learning materials). They show that the most likely explanation for this result is that household spending on books and school supplies did not change across treatment and schools in the first year (when the school grant was unanticipated), but that households in treatment schools sharply cut back their spending on these categories in the second year (when the school grant was anticipated and could be accounted for in household decision making), and that this reduction offset around 80% of the per-student value of the grant.

The authors therefore argue that the “first-year” effect of the program is more likely to represent the “production-function” effect of providing the school grant (since other factors did not have time to adjust), whereas the “second-year” effect is closer to the “policy parameter” (which reflects household reoptimization). The example highlights the value of measuring as many intermediate inputs as possible to have a better idea about the mechanisms of program impact. However, in practice, it will be difficult to measure *all* possible intermediate inputs, and the extent to which they may have changed in response to the exogenously varied treatment. Thus, it is perhaps most accurate to interpret the “causal estimate” of  $\beta$  from experimental studies as the “policy effect” of  $X_{i,t}$  at the point when the outcomes are measured.

Note that this limitation is also present for nonexperimental methods, and is therefore not a criticism of experiments per se. But it is an important limitation to highlight because experimental estimates are often implicitly interpreted as representing production function parameters based on Eq. (1). This may well be true over short time periods where other agents may not have reoptimized behavior, but it is (1) difficult to confirm that this is true on every dimension of potential behavior modification, and (2) much less likely to be true over longer horizons.<sup>16</sup> One advantage of well-identified evaluations using large administrative data sets (based on regression discontinuity designs, for example) is that it may be possible to observe the policy effects at longer time horizons at much lower

<sup>16</sup> While the discussion may suggest that experimental estimates may be lower bounds of production–function parameters and upper bounds of policy parameters, this need not be true if the unmeasured inputs are complements to the experimental intervention as opposed to substitutes (as was the case in Das et al., 2013).

marginal cost than in experimental studies (since the cost of conducting follow-up surveys on experimental samples can be quite large, and the challenge of differential attrition grows over time). A good example is provided by [Bharadwaj et al. \(2013\)](#) who can measure the impact of early childhood interventions several years later using administrative data in both Chile and Norway. Longer term follow-ups of experimental interventions are relatively rare, but should be a higher priority for funders and researchers.

This discussion also unifies the approaches outlined in [Sections 2.3 and 2.4](#). One advantage of using the approach in [Attanasio \(2015\)](#) (as outlined in [Section 2.4](#)), which evaluates experiments through the lens of a constrained household choice problem is that reoptimizing behavior is directly built into the problem framework as opposed to being an afterthought. Therefore, it forces the researcher to be disciplined about how the intervention affects either the constraints (production function, resources), beliefs/attention (as with knowledge interventions), or household preferences (as with interventions that may affect intrahousehold bargaining) and to interpret the effects through this unified lens.

## 4.2 Interpreting zero effects

A second challenge in conducting inference from experimental studies is interpreting zero effects. In theory, this should simply mean that the estimate of  $\beta$  in Eq. (1) is zero and that the marginal impact of increasing  $X_{i,t}$  is insignificantly different from zero. In practice, however, it is important to distinguish between five different possible interpretations of a zero result. These include (1) poor implementation of the intervention, including corruption or administrative failures, (2) substitution away of other inputs by agents (including governments, schools, teachers, and households) in response to the treatment, (3) positive effects on some subpopulations but not on others, leading to an average effect that is not significantly different from zero, (4) absence of complementary inputs/reforms that may be needed for the intervention to be effective, and (5) a true zero effect for all students. Note that reasons (3), (4), and (5) are consistent with the interpretation that the marginal impact of increasing  $X_{i,t}$  on outcomes is zero in a production function sense, but reasons (1) and (2) are not. Further, the distinction between (3), (4), and (5) also matters because the policy implication of (5) would be to not prioritize increasing  $X_{i,t}$ , whereas that of (3) would be to provide it to the subgroup where it was effective, and that of (4) would be to increase  $X_{i,t}$  as long as the complementary input is also increased.

These possibilities are illustrated across four different randomized evaluations of the impact of providing books and materials to students. Each of the four studies find zero-average impacts of providing books and materials, but point to different possible reasons for the zero effects. [Sabarwal et al. \(2014\)](#) find no impact on test scores from the provision of textbooks to schools in Sierra Leone and attribute this to the fact that

schools actually stored the textbooks instead of distributing them to students (which is a form of poor implementation). Das et al. (2013) described above also find no net impact on test scores from the provision of a school grant (that was mostly spent on books and materials) in India, but attribute it to households offsetting the intervention by reducing their own spending on these inputs. Glewwe et al. (2009) also find no impact on test scores from providing textbooks to students in Kenya. However, they find positive impacts on students with the highest baseline test scores and suggest that their results are consistent with the fact that the majority of children could not read the English language text books to begin with; thus, they could not benefit from the textbooks (whereas those who could read *did* benefit).

Finally, Mbiti et al. (2016) also find no impact on test scores from the provision of a large capitation grant to schools in Tanzania (the largest item that the grant was spent on was textbooks). However, their study was explicitly designed to test for complementarities with teacher effort (which was boosted by a separate intervention that paid teachers bonuses based on student performance) using a cross-cutting design with a sample size large enough to test for complementarities, and they find that the interaction effect of the school grant and teacher-performance pay was significantly positive. In other words, the school grant on its own had no impact, but had a significant impact when provided in conjunction with a teacher-performance-pay intervention. Thus, it is likely that the performance-pay treatment contributed to teachers making more effective use of the additional materials, but it is also true that having the materials allowed teachers to significantly improve student outcomes relative to teachers who only increased effort due to performance-linked pay.

The larger point here is that each of these experiments with zero results are useful results in and of themselves, and yield an important policy conclusion that the marginal impact of providing books and learning materials to students may be very low on their own. On the other hand, the fact that four papers with the same result point to four different reasons for this nonimpact suggest that a “black-box” experiment on its own may yield limited insights into the nature of the education-production function and the true binding constraints to learning.

More generally, it is important and useful to document *why* interventions that funders and implementers spend so much time and money on have no impact (if they do not). Papers that find zero impact are very important, but will typically contribute more to learning if accompanied by careful analysis of intermediate variables to better understand and describe which parts of the posited causal chain of impact worked and which ones broke down. Thus, it is a good practice for researchers to think *ex ante* about how they would interpret a zero effect, and to collect as much data as possible on implementation quality as well as intermediate inputs and processes to enable better interpretation of finding no effects of a program.

## 4.3 External validity

Perhaps the most widely discussed limitation of experiments is the external validity of their results beyond the specific setting where they are carried out (Cartwright, 2007; Deaton, 2010). The formal way of thinking about this problem is to recognize that though the random assignment ensures that unobservables are distributed identically across treatment and control groups and that the treatment is not correlated with these unobservables, the estimated program effects are for not for the treatment *alone*, but rather for the treatment *interacted* with the unobservable characteristics in the study sample. If these unobservable characteristics vary between the study sample and the universe to which we seek to extrapolate the findings to, then the estimated treatment effects may not be valid because the interactions may change.

There are several variants of this concern that are worth spelling out distinctly, because the strategies for mitigating them are different. There are at least four limitations to generalizing experimental results even in the *same context* where the experiment was carried out. I discuss these first before discussing external validity *across contexts*.

### 4.3.1 External validity in the same context: representativeness of study samples

First, there is a concern of external validity even in the context of the evaluation because most experiments are carried out within a universe of schools that agree to participate in the experiment. If these schools are different from those who do not agree to participate (perhaps their leadership is more open to trying out new ideas), then the results might have limited external validity (Heckman and Smith, 1995).<sup>17</sup> Most experimental studies in education do not pay enough attention to this issue, and it is not that difficult to do so (see Section 5.2.1).

### 4.3.2 External validity in the same context: implementer heterogeneity

Second, a further concern with external validity even in the same context comes from the fact that many RCTs in education evaluate interventions implemented by committed implementation partners [often highly motivated non-governmental organizations (NGOs)]. Thus, experimental estimates of programs implemented by NGOs may not generalize if the same program is implemented by the government (as shown in Bold et al., 2013). The differences in the results they report between government and NGO implementation of a contract-teacher program largely reflects the fact that the program itself was poorly implemented by the government. So, it does not negate the results found under NGO-implementation, but it does highlight that programs

<sup>17</sup> A variant of this concern is seen in the US charter-school literature where well-identified estimates are only available on the causal impact of oversubscribed charter schools (which are likely to be the higher quality ones) as opposed to the universe of charter schools, which is the policy parameter of interest (unless the oversubscribed schools are able to expand without compromising quality and the schools that are not oversubscribed shut down).

are not just an “intervention”, but rather an intervention *and* an implementation protocol. This is not a problem per se, but suggests that evaluations of NGO-led implementations should be seen as efficacy trials and not effectiveness trials.<sup>18</sup> It also suggests that when successful NGO-implemented interventions are being scaled up, there may be a strong case for conducting further RCTs at larger units of implementation and when implemented by the entity that will eventually scale up the intervention (typically a government).

#### **4.3.3 External validity in the same context: varying intervention details**

Third, even if a government wishes to use experimental results from a given context to guide policy in the same context, it is unlikely that the policy chosen will be exactly the same as the one evaluated. For instance, even if a CCT is found to have a positive impact, the value of the CCT may be changed later for political or budgetary reasons. Similarly, even if a teacher performance-pay program or a student-incentive program is found to be effective, a policy maker would care about the elasticity of the outcome of interest to the magnitude of the incentives in order to better calibrate the value of the incentives. These questions are harder to answer within the context of an experiment because it is often politically and administratively difficult to vary the magnitudes of such incentives within an actual program. Further, experimentally estimating elasticities often requires sacrificing power or a larger sample. While it may be possible to do this, a more promising approach may be to combine experimental methods with structural modeling to allow more credible out of sample predictions than either of the two approaches could on their own.

Good examples of this approach are [Todd and Wolpin \(2006\)](#) and [Attanasio et al. \(2012\)](#) who combine structural models of school participation with observed impacts of the PROGRESA CCT program to enable predictions of program impact under alternative values of the cash transfer. Another good example is [Duflo et al. \(2012\)](#) who use the nonlinearities in the compensation schedule of an experimentally evaluated teacher incentive program to identify parameters in a dynamic model of teacher-labor supply and use the model to estimate cost-minimizing compensation policies to achieve a desired level of teacher attendance. However, these additions of structural models to enable out of sample predictions have mostly been done ex post and were not designed ex ante into the study, which may have limited the extent to which the experiment could be used to identify parameters in the structural model of interest. Future experimental research in education is likely to have greater impact if the ex-ante design of the study

<sup>18</sup> These terms are standard in the medical literature and refer to the difference between impacts under high-quality implementation that is closely monitored (efficacy trial), and impacts under typical implementation that allows for typical patient behavior including noncompliance with dosage frequency and complementary instructions (effectiveness trial).

includes careful thinking about the model that the experiment can be used to identify, and plans its data collection strategy accordingly.<sup>19</sup>

In practice, the skill set required to run high-quality field experiments is considerably different from that required to specify and estimate structural models. Thus, there are likely to be considerable benefits to forming teams of researchers with complementary skills across designing and running field experiments and structural modeling from the outset to conceptualize experiments from the dual perspectives outlined in Sections 2.3 and 2.4 and to design data collection strategies accordingly.

#### **4.3.4 External validity in the same context: political economy**

A fourth and final concern regarding external validity in the same context is that experiments cannot typically capture the general equilibrium effects (both political and economic) that may accompany attempts to scale up successful smaller scale experiments. In the words of Acemoglu (2010), “Political economy refers to the fact that the feasible set of interventions is often determined by political factors, and large counterfactuals will induce political responses from various actors and interest groups. General equilibrium and political economy considerations are important because partial equilibrium estimates that ignore responses from both sources will not give the appropriate answer to counterfactual exercises”.

A good example of this is the case of contract teachers, where existing experimental and nonexperimental evidence suggest that locally hired contract teachers who are less educated, less trained, and paid much lower salaries than civil-service teachers are at least as effective (if not more) at improving learning outcomes in rural primary schools in both Kenya and India (Duflo et al., 2015b; Bold et al., 2013; Muralidharan and Sundararaman, 2013). Thus, expanding the use of contract teachers on the current margin would appear to be a very promising and cost-effective policy for improving education outcomes in developing countries. Nevertheless, scaling up contract-teacher programs has been difficult politically because forward looking officials are aware that hiring a large number of contract teachers will lead to them getting unionized and creating political pressure to get “regularized” as civil-service teachers, which is very difficult for politicians to ignore.

<sup>19</sup> For instance, household surveys that accompany school-level treatments are often collected in repeated cross-section samples. This makes sense if the goal is to characterize the *average* difference in household inputs across treatment and control groups, because cross-sectional heterogeneity is typically much greater than over-time heterogeneity within the same household, and a repeated cross-section that covers more households will yield more precise estimates of these averages. However, a structural modeling exercise that aims to understand household decision-making in response to a school-level treatment will typically benefit from repeated surveys of a smaller sample of households to better model the dynamics of household choice in response to realizations of information from their child’s school over time.

This is a problem that is difficult to solve empirically with an experiment, but the discussion above highlights the importance of treating positive results from an experimental evaluation of an intervention as just one of many inputs into the policy-making process. Finding positive technical results from an RCT of an intervention can be a good starting point for considering the administrative and political challenges of scaling up and designing implementation protocols that take these into account, but it would be naïve to recommend “scale-ups” based on RCT evidence alone. It is perhaps not a coincidence that the leading example of policy scale-up based on RCT evidence is deworming, which is administratively easy and politically costless. In contrast, other interventions with robust evidence of positive effects (like the use of contract teachers) have been much more difficult to scale up. In such cases, the experimental evidence is best treated as the starting point for a more informed policy conversation.<sup>20</sup>

#### **4.3.5 External validity concerns across contexts**

Obtaining external validity across contexts is even more challenging, given that the unobserved covariates (that would interact with the treatment of interest to produce the average treatment effect) are likely to be different across contexts. This problem is well known among academic researchers, but is often understated in “systemic reviews” that compare interventions and estimates across contexts (see [Pritchett and Sandefur, 2013](#) for a discussion). There is no good experimental solution to this problem beyond conducting more studies and gathering more evidence by replicating evaluations of similar (if not “identical”) interventions in many settings. The problem of external validity from well-identified individual studies is now receiving more formal attention (see [Dehejia et al., 2015](#) for an approach that derives an external validity function based on matching on observed covariates), and is likely to be an area of fruitful research because the analytical standards that have been applied to external validity in the past decade have been much weaker than those applied to internal validity.

Despite the many successful field experiments in education in developing countries in the past decade, the overall experimental research agenda in this area is still at an early stage. Some pieces of evidence seem quite robust across several contexts (such as the lack of impact of providing books and materials to students), and others have been replicated in multiple sites in the same country (such as “TaRL” across states in India), but most other interventions do not have enough replications across contexts to enable confident claims of their impacts across contexts. A further problem is that there is considerable variation in the details of interventions implemented and evaluated across settings, which

<sup>20</sup> See [Muralidharan \(2013, 2016\)](#) for an example of a policy proposal that takes the results from evaluations of contract teacher programs seriously, and accounts for administrative and political economy considerations in recommending feasible policy approaches that are consistent with the evidence.

makes generalization even more challenging (see the discussion in [Glewwe and Muralidharan, 2016](#) on this).

Thus, to the extent that donors and development agencies seek summaries of evidence (as seen by the several summaries written in the last few years), attempts to calculate comparative cost effectiveness of interventions conducted across several contexts should be interpreted cautiously ([Dhaliwal et al., 2012](#)). In trying to learn across contexts from the experimental literature to date, it may be more appropriate to focus on “principles” that have been validated in multiple settings rather than the “point estimates” of specific studies. The summary of the evidence presented in [Section 3](#) reflects this approach.

## 5. CONDUCTING FIELD EXPERIMENTS IN EDUCATION IN DEVELOPING COUNTRIES

There are several excellent resources for researchers and practitioners wanting to design, implement, and analyze field experiments including [Duflo et al. \(2007\)](#), [Gerber and Green \(2012\)](#) and [Glennerster and Takavarasha \(2013\)](#). Nevertheless, there is a considerable amount of tacit knowledge that is accumulated by researchers through practice in a particular area (such as education) that is often not available easily, and the goal of this section is to synthesize some of this knowledge and offer it as a resource for researchers and practitioners wanting to design, conduct, analyze, and interpret field experiments in education in developing countries.<sup>21</sup>

### 5.1 Design

The core of a good experiment is its design, which requires considerable *upfront* thought and attention to develop. The importance of investing in this upfront thinking is perhaps best illustrated with an anecdote about doing economics research in the 1950’s that I heard from a very senior researcher. He mentioned how a graduate student doing empirical work would often just be able to run *one* regression for an entire dissertation since computer time would have to be booked many months in advance and punch cards carefully prepared in advance of being able to run the one regression. Thus, all the thinking had to be done in advance, and had to be checked and re-checked multiple times because you would lose months of time if there were errors in your code or data.

Conducting field experiments is similar in that you typically only get to run the experiment once, and so it is essential to get all the thinking done upfront. This may

<sup>21</sup> There is nevertheless a high degree of “learning by doing” in conducting field experiments. It is therefore a common practice (and highly advisable) for young researchers to initially work on teams led by senior researchers (including as research assistants) to obtain the tacit knowledge that is best obtained by practice.

seem obvious, but the best experimental papers have a deceptive “easiness” about them that typically hides the large amount of advance thinking that goes into a well-done experiment. Experiments are unforgiving of mistakes as it is typically not possible to change the design of treatments or the overall experiment once it is underway. Thus, it is essential to obtain multiple sources of feedback before commencing a field experiment, and the best experimental designs are often iterated several times at the design stage. The discussion below covers some of the main considerations in designing a good field experiment in education and is organized around two main topics — intervention design and experiment design.

### **5.1.1 Intervention design: what to look for?**

While experimental methods can help with credible estimation of the causal impact of interventions, it is important to ensure that adequate thought is given to determining whether the intervention being studied is worth evaluating and the extent to which the findings of a study generate more generalizable knowledge (especially given the nontrivial time and effort costs of setting up an RCT). In particular, it is not uncommon to see competently implemented and analyzed experiments in education, where the underlying intervention is rather ad hoc and not adequately theorized, which limits what we can learn from the evaluation. In this section, I offer some (personal) guidelines for informing the decision on whether an intervention is worth evaluating experimentally.

Three useful questions to ask before deciding whether or not to conduct an experimental evaluation of an education intervention are: First, is there genuine uncertainty about the impact of the intervention. Second, is the intervention addressing a well-understood supply or demand-side deficiency and *designed well enough* to address this deficiency. Third, are governments spending large amounts of money doing things whose effectiveness we do not understand well enough (e.g. infrastructure, class size reductions, teacher training, teacher salary increases, school feeding programs, school grants)?

A good rule of thumb for identifying whether an intervention is worth studying experimentally is not just to want to test “if a program works” but to test ideas where there is genuine uncertainty and controversy regarding their impact. Experimental evaluations of education interventions are typically more influential when there are compelling theoretical arguments both in favor of and against an intervention, and where the answer is essentially an empirical question. Examples include the impacts of student tracking ([Duflo et al., 2011](#)), the impacts of linking teacher pay to measures of effort ([Duflo et al., 2012](#)) or gains in student test scores ([Muralidharan and Sundararaman, 2011](#)), and the impacts of school choice and private schools

(Muralidharan and Sundararaman, 2015). These studies are characterized not just by studying the impact of the intervention on education outcomes, but by paying serious attention to the hypothesized negative impacts and taking care to measure these potential negative impacts to reach a more nuanced and complete understanding of the impact of the intervention.

In the case of tracking, the concern of opponents of tracking regarding negative peer effects on students assigned to the “lower performing” track was tested by Duflo et al. (2011) by combining an experimental evaluation of tracking with a regression-discontinuity-based evaluation of whether there were negative peer effects. In the case of the teacher performance-pay program evaluated by Muralidharan and Sundararaman (2011), the authors designed the program to mitigate against some of the known concerns of teacher performance-pay programs and tested for others (such as teaching to the test and potential negative effects on nonincentive subjects). In the case of school choice, a key concern has been the possibility of negative spillovers on students who did not apply for the voucher or who were in private schools to begin with and were exposed to lower achieving peers who transferred in to these private schools with vouchers. The study conducted by Muralidharan and Sundararaman (2015) addresses these issues with a two-stage randomization design that allowed the researchers to quantify these spillovers.

A second class of interventions worth studying are those where there is a well-hypothesized theory of change from an intervention to outcomes and the program is not necessarily controversial, but where the program is not being implemented (typically due to the lack of a champion within the government or budgetary constraints). In these cases, a high-quality evaluation accompanied by cost-effectiveness calculations can be a very useful contribution to research and policy by helping to make the case for allocating public funds for expanding the program (if effective) or for not doing so (if found ineffective).

A good example is the provision of deworming tablets to school children, which was found to be a much more cost-effective way of increasing school attendance than other spending on other student inputs (Miguel and Kremer, 2004; Dhaliwal et al., 2012), and has since resulted in scaled-up deworming programs in many parts of the world. A second example is the growing body of evidence on the effectiveness of “TaRL” (see Banerjee et al., 2016), which is increasingly leading to governments being interested in scaling up the core idea.

Both these ideas seem obvious ex post, but did not receive much policy maker attention before the research results. In the case of deworming, the cost of the intervention was trivial, but school health programs (under which such an initiative would typically have to be implemented) would often suffer from coordination failures across health and education ministries, without clear ownership. In the case of “TaRL”, the idea

that children are not learning because they are far behind where the default level of classroom instruction simply does not seem to cross the mind of many policy-making elites because the situation is outside most of their own experiences.<sup>22</sup> Thus, high-quality evaluations of innovative interventions can be a useful catalyst in making nonobvious constraints to education quality salient to policy makers, and obtaining policy consideration for cost-effective solutions to alleviate these constraints.

A third category of evaluations worth conducting are those studying policies or programs that governments spend a lot of money on. Even if the researcher has *ex ante* reasons to believe that the intervention is not well designed or may not be effective, these evaluations can be very useful because the money will be spent on the program anyway, and it is important to understand whether the program was effective. Further, to the extent that the program (as designed) reflected conventional wisdom that it would have a positive impact, the evaluation could shed light not just on the “program” as implemented but also on the hypotheses underlying the design of the program.

A good example of such an evaluation is [de Ree et al. \(2015\)](#) who study the impact of an unconditional doubling of teacher pay in Indonesia on the performance of incumbent teachers. While economists may have had a prior expectation that this doubling may not have much impact on student learning (or at least that the same money could have been much better spent), the evaluation was still worth conducting because (1) the policy was very expensive, and (2) many education advocates believed that increasing teacher pay would improve teacher motivation, effort, and student learning.

### **5.1.2 Intervention design: what to avoid?**

One common mistake is to rush into an RCT before the intervention being studied has been adequately piloted, codified, and stabilized. If the intervention is modified during the study, it is difficult to interpret the findings. Thus, it is advisable for programs and interventions to be “standardized” and “easily replicable” before embarking on an RCT. A related challenge occurs with RCTs of “composite” interventions that include components that are not easy to codify, which makes it difficult to interpret the results of an RCT. Note that a “composite” intervention *per se* is not a problem since there are

<sup>22</sup> Indeed, one plausible reason for why well-intentioned education interventions may have limited impact on learning outcomes is that policies and programs are typically designed by elites whose experiential understanding of education may not correspond to that of the representative student. For instance, on a field visit to schools in Tanzania, I saw that several children would typically share one textbook. It is not difficult to imagine how this would be a highly salient fact for a visiting senior education policy official or foreign aid official, and result in a well-intentioned program to provide free textbooks that would take considerable financial and administrative resources to deliver. However, as the results in [Glewwe et al. \(2009\)](#) suggest, the provision of free textbooks may not alleviate the binding constraint for the average student in this setting, which is that they cannot read. Similarly, media discussions of education in countries such as India focus disproportionately on the issues relevant to the high end of the achievement distribution (who comprise the readership of newspapers) as opposed to issues relevant to the representative student.

often complementarities across components of the composite “package” intervention, and the “package” may be the intervention that we need to evaluate. Rather, it is the inability to codify and replicate the “whole package” that limits the learning from an evaluation of composite interventions.

Another pitfall to be aware of is studying “gold-plated” interventions that have high unit costs. In such a setting, a high-quality (but high-cost) intervention may get evaluated and found to have a substantial positive impact, but it may be difficult to scale up because of a lack of financial resources to sustain the program.<sup>23</sup> A further risk is that a diluted version of the high-cost prototype is scaled up (on the basis of the evaluation), but that this version may not have any impact. One option for imposing discipline in this regard is to not only have a “pure” control group (that does not get any additional intervention), but to have other comparison groups that are provided an equivalent amount of resources, which enables a direct cost-effectiveness comparison against reasonable policy alternatives. An example of this is provided by the Andhra Pradesh Randomized Evaluation Studies (APRESt) where the impact of teacher-performance pay was evaluated not just against a pure control group, but against comparison groups that received an equivalent valued school grant or extra contract teacher.<sup>24</sup>

Finally, one should be wary of evaluations of policies and programs whose details have not been well designed, in which case even an experimental evaluation may not contribute much to learning. This point is best illustrated through the example of teacher-performance pay policies. Typically, these policies are created when a policy maker decides to implement some form of teacher-performance pay, and administrators then design a particular formula for paying teachers performance-based bonuses. However, the optimal design of a performance-pay system is a nontrivial problem, and in many cases, the formulae designed by administrators are likely to have important design flaws that are likely to limit the effectiveness of performance pay. For instance, many simple formulae reward teachers based on the number of students who pass a performance threshold (like passing a test), which does not provide incentives for teachers to continue improving student learning above the threshold, or far below the threshold (see [Neal and Schanzenbach, 2010](#)). In such a case, an evaluation that does not find a significant impact of performance pay is not very useful beyond being able to say that the specific program as implemented was not effective.

<sup>23</sup> One manifestation of this is the phenomenon of donor-financed pilot projects being abandoned once the donor financing is over. Of course, these programs typically do not have credible impact evaluations to inform the decision on whether the developing-country governments should continue to spend on them out of their own budgets. But scaling up of high unit-cost interventions would be difficult even with positive evidence of impact.

<sup>24</sup> [Blattman and Niehaus \(2014\)](#) make a similar point with regard to evaluations of antipoverty programs in general, proposing that the benchmark should not just be a pure control group, but rather an unconditional cash transfer that is equal in value to the full cost of the program being evaluated.

The design of teacher- and student incentives is a good example of a case where economists can add value not just in terms of conducting a “well-identified evaluation”, but in using theoretical first principles to design better variants of the intervention. There is an extensive theoretical literature on incentive design that covers topics ranging from: piece rates versus tournaments ([Lazear and Rosen, 1981](#); [Green and Stokey, 1983](#)); linear versus nonlinear incentives ([Holmstrom and Milgrom, 1987](#)); group versus individual incentives ([Itoh, 1991](#); [Kandel and Lazear, 1992](#); [Kandori, 1992](#)), and a corresponding empirical literature with examples from outside education. There is also a theoretical literature on optimal design of teacher-performance pay systems (see [Barlevy and Neal, 2012](#)) that has implications for the optimal design of student-incentive programs. There are several papers on teacher/student incentives that do not pay attention to this theoretical literature and are weaker as a result.

To summarize, it is important for researchers to pay at least as much attention to the *design of the intervention* as to the design of the evaluation. Further, when researchers have influence over the design of the intervention, they should aim to understand the relevant theoretical literature to inform the decisions and trade-offs that will have to be made in finalizing the intervention. More generally, any field experiment in education will benefit from researchers starting the study by clarifying the research question, and asking themselves what they would learn about the world and the intervention from different potential values of the treatment effect.

### **5.1.3 Experiment design: unit of randomization**

As discussed earlier, the “modular” nature of many education interventions makes it feasible to randomize them at relatively small unit levels, including at the student, classroom, and school levels. The main consideration in determining the unit of randomization is the trade-off between statistical power<sup>25</sup> for a given budget (which is higher for lower units of randomization) and the possibility of spillovers, which may bias the experiment and negate the power gains from randomizing at a lower level.

A striking example of this trade-off is seen in the difference between impact evaluations of deworming that randomized the treatment at the student level and those that randomized the treatment at the school level, such as [Miguel and Kremer \(2004\)](#). While the former typically found very limited impacts of deworming on the education outcomes, [Miguel and Kremer \(2004\)](#) found large positive impacts on treated students, as well as spillovers from treated to nontreated students. The spillovers would underestimate the impact of deworming in studies with student-level randomization in two ways. First, it would underestimate the impact on the treated students (because the nontreated

<sup>25</sup> Note that this chapter does not spend too much time on generic issues of experimental design such as power calculations, for which there are many other references. Rather the focus is on design choices that are especially relevant to research in education.

students who served as the control group also benefited from the intervention). Second, it would not count the impact on nontreated students who also benefited from the intervention. Of course, randomizing at the school level significantly increases the sample size required for adequate power, and correspondingly increases the cost of the study. On the other hand, it is not clear that the gain in power from student-level randomization is worth it if it biases the treatment effect itself.

Note, however, that the case of deworming may be an outlier in terms of the extent of spillovers across students. Several studies within the “Rural Education Action Program” (REAP) have utilized student-level randomization within schools in China to study the impacts of interventions ranging from peer-tutoring to student incentives. Another example is provided by [Berry \(2015\)](#), who uses student-level randomization to study the impact of different combinations of student and parent incentives.

Nevertheless, given that students interact with each other every day in the classroom, it is difficult to credibly claim that there would not be any spillovers (especially for treatments implemented in schools as opposed to households), which may limit the extent to which we can learn from such experiments. One important exception is cases where the treatment is at the school level and control-students are not in treated schools, as happens in cases where students receive vouchers (or charter-school admission) by lottery and do not interact with control students during the course of the school day. But overall in general, student-level randomization designs should be used with caution and limited to situations where the focus of the intervention is outside the school setting (such as households or after-school programs).

A less extreme set of concerns applies to designs that randomize at the grade or classroom level as opposed to the school level. Again, the main reason for doing so is power and cost. Several prominent studies have used classroom-level randomization designs to study the impact of remedial instruction ([Banerjee et al., 2007](#)), tracking of students according to initial ability ([Duflo et al., 2011](#)), and comparing contract and regular teachers ([Duflo et al., 2015a](#)). The spillover concerns are less severe in this case because most instructional activity as well as peer interactions between students happens at the classroom level and not across classrooms (which is where the spillovers would have to happen).

However, there is still a concern that interventions that provide a significant increase in resources to some classrooms may lead a head-teacher to offset some of the impact of the treatment by reallocating some other resources to control classrooms. This could happen either due to norms of fairness within the school or because an optimizing head-teacher would reallocate resources to equalize their marginal product across classrooms.<sup>26</sup> Such behavioral responses could contaminate the experiment and the inference.

<sup>26</sup> For instance, the model in [Lazear \(2006\)](#) predicts that more disruptive students will optimally be assigned to smaller classrooms.

This is why my personal preference has been to randomize at the school level to the extent possible for studying interventions ranging from teacher-performance pay, across the board teacher salary increases, provision of school grants, provision of diagnostic feedback to schools, and also the provision of an extra contract teacher. The logic of this approach is that a policy maker can typically target resources at the school level, but cannot easily control how those resources are allocated within the school by an optimizing head-teacher. Thus, the policy-relevant parameter of interest in these cases is the impact of a school-level provision of an intervention. The limitation of this approach of course is that the samples need to be much larger and the studies cost more.

One rule of thumb for choosing between school and grade/classroom level randomization is to consider the size of the school. In smaller schools (as is the case in rural India where most primary schools have less than 100 students across 5 grades and the modal school has only two teachers teaching in multigrade classroom settings), it is difficult to convincingly argue that a within-school randomization protocol that assigns a program to just some grades will be adhered to without readjustment. On the other hand, when schools are much larger and have several hundred students and dedicated teachers in each grade (as is the case in many African settings), the fidelity of within-school experimental protocols may be more reasonable to assume because teachers spend all their time with one grade as opposed to teaching multiple grades at the same time. Further, the costs of school-level randomization may be prohibitive in settings with such large schools.

Overall, I believe that it is less problematic for an intervention to be targeted at specific grades within a school (as opposed to the entire school) as long as the control group is the same grade in a *different* school as opposed to other students within the same school. In such a setting, one may still worry about spillovers attenuating the treatment (if resources are diverted to other nontreated grades in treatment schools), but at least the control group will not be contaminated. Studies that use within-school controls should have the burden of proof for demonstrating that the controls were not contaminated by spillovers by the time that the end-line measurements are conducted.

#### **5.1.4 Experiment design: power and baselines**

Beyond standard discussions of power calculations and sample size calculations, the following three strategies can increase power in education interventions. First, given the autocorrelation in student test scores over time, it usually makes sense to conduct student-level baseline tests and include these test scores as controls in estimates of treatment effects. Second, additional gains in precision (and power) can be obtained by also controlling for school-grade-subject level baseline means in the treatment-effect regressions, in addition to the student's own lagged test scores ([Altonji and Mansfield, 2014](#)). Finally,

stratifying the randomization within geographic/administrative regions can further increase power ([Muralidharan and Sundararaman, 2011](#); [de Ree et al., 2015](#)).

Nevertheless, it can also make sense to not do a baseline and it is possible to conduct high-quality experimental studies without a baseline round of testing. Reasons to not do a baseline include: risk management, time, and budget. In settings where there is a reasonable risk that the intervention may not be implemented, it may be prudent to not do a baseline, conduct a randomization for the intervention roll-out, and only expend time and effort on an evaluation if the intervention was actually implemented, and the randomization was successfully adhered to. Such an approach may also be needed in cases where there isn't adequate time to put in place field teams and raise research funding for the baseline. In such cases, it may be adequate to use administrative data for conducting the randomization (even if the administrative data is only at the school level and not the student level).

This approach may be especially useful when researchers are working with governments, where the risk of nonimplementation of the intervention and noncompliance with a randomization protocol are higher. In such cases, it may make sense for researchers (especially risk-averse junior researchers with tighter resource constraints) to initially focus efforts on ensuring randomization and implementation of the intervention (and to push for a larger scale of implementation if possible), and to compensate for the lack of a baseline by increasing the sample size of the endline. Note that under government implementation, the sample size (and power) is usually not constrained by the intervention budget but by the evaluation budget and sample.

Finally, another reason for young researchers to consider this approach is that funders are more likely to view your proposal favorably if you can demonstrate that the implementing agency has demonstrably adhered to the randomization and experimental roll out protocol. So an optimal approach may be to apply for smaller amounts of pilot funding to travel to the location of the study, interact with implementation partners, influence the intervention design (to the extent possible), collect administrative data, conduct the randomization, and ensure that the randomization protocols are followed during implementation. This can then be followed up by applying for funding for the endline, at which stage the implementation risks of the project would have been lowered considerably.

This is also a good place to discuss the issue of cost effectiveness of study designs. While many of the points made in this section may seem like they can only be addressed by senior researchers with larger budgets, the discussion above highlights that it is often possible to considerably reduce the cost of an RCT. In particular, studies that use administrative data on student outcomes can be especially cost-effective to conduct, and should be a high priority for both researchers and funders going forward.

### **5.1.5 Experiment design: cross-cutting designs and interactions**

Many experimental studies on education in developing countries aim to compare the relative effectiveness of different interventions on education outcomes in the same setting. One commonly used approach to reducing the cost of conducting multiple studies in a given setting is to employ “cross-cutting” or “factorial” designs. In its simplest form, this usually involves a two-by-two matrix of four treatment cells with one group receiving neither treatment (the control group), a second group receiving just the first treatment (T1), a third receiving just the second treatment (T2), and a fourth receiving both (T1 and T2). The logic of such cross-cutting designs is best expressed by Kremer (2003) who notes that: “Conducting a series of evaluations in the same area allows considerable cost savings. Since data collection is the most costly element of these evaluations, cross-cutting the sample reduces costs dramatically. *This tactic can be problematic, however, if there are significant interactions across programs.*”

Several highly influential papers have taken this approach to randomization and a particularly good example is Duflo et al. (2011) that was conducted as part of an “Extra-Teacher Project” that aimed to study several important questions simultaneously in the same setting including (1) the impact of class size reduction, (2) the impact of contract versus regular teachers, (3) the impact of tracking students by initial test scores, and (4) the impact of community monitoring of schools. The project managed to study all four of these questions with a sample of just 210 schools by using a cross-cutting design with 70 control schools and 140 schools assigned to a combination of the treatments. The key to generating consistent estimates with this approach is to precommit to a view that interactions among treatments are not important. This allows the researchers to treat the treatment effects as additive and makes it possible to double-count the schools with multiple treatments under each treatment (to increase power within a given measurement budget).

But in practice, interactions *are* likely to be important, and there is a risk of obtaining biased estimates of the impact of individual treatments if there are interactions across treatments. Mbiti et al. (2016) show this in the context of an experimental evaluation of school grants and teacher incentives in Tanzania, that was explicitly powered to test for complementarities (and committed to this in a preanalysis plan). They show that (1) each treatment was insignificant on its own when interactions are accounted for, (2) that the interactions between treatments were positive and significant, and that (3) ignoring the interactions would lead to an overstatement of the individual-treatment effects.

Thus, the concerns about cross-cutting designs expressed in Kremer (2003) are likely to be salient in practice. The quote from Kremer (2003) highlights that this was and is a well-known issue. But it made sense for early experiments to use cross-cutting designs

and ignore interactions at a time when evaluation budgets were much tighter, because this allowed adequate power for the first order questions.

However, as we generate evidence that interactions do matter (and are detectable with designs that treat them as equally important as the main effect and allocate adequate sample size to detect them), the assumptions underlying cross-cutting designs that ignore interactions appear less tenable. Since evaluation budgets are growing (with programs increasingly being expected to set aside funds for evaluation), it may make sense to prefer cleaner designs with single treatments and direct comparisons of treatments without being confounded by interactions. It may still make sense to ignore interactions in some settings, but doing so should be justified and documented in a preanalysis plan. Note that this does not imply that every component of treatments should be broken down and tested separately. Many treatments of interest are “composite” by design and should be evaluated that way. But such an approach does not assume that the interactions are zero, which is what cross-cutting designs do (see [Mbiti et al., 2016](#) for a more extended discussion on the implications for experiment design and hypothesis testing).

## 5.2 Sampling, randomization, and implementation

### 5.2.1 Sampling and representativeness

As mentioned in [Section 4.3.1](#), experimental papers in education often do not pay enough attention to the representativeness of the universe of schools, which can limit the external validity of the studies even within the context in which they are conducted. The ideal experimental protocol to address this problem is to try as hard as possible to first draw a representative sample of schools/students from the universe that the study is trying to extrapolate to, and then randomly assign these schools into treatment and control groups. Such a protocol provides much more external validity than studies carried out in a “convenience sample” of schools and allows policy makers to be more confident that the experimental estimates apply to the relevant universe of interest. Examples of such an approach include [Muralidharan and Sundararaman \(2011, 2013\)](#) in the Indian state of Andhra Pradesh, [de Ree et al. \(2015\)](#) in Indonesia, and [Mbiti et al. \(2016\)](#) in Tanzania. Each of these studies featured random assignment in near-representative samples that allowed the results to be credibly extrapolated to populations of  $\sim 80M$  (Andhra Pradesh),  $\sim 200M$  (Indonesia), and  $\sim 45M$  (Tanzania).

While this may not always be possible for reasons of cost and logistics, experimental studies should at least discuss their sampling procedure in more detail (which is often not done) and show tables comparing the study sample and the universe of interest on key observable characteristics (similar to tables showing balance on

observable characteristics across treatment and control units). Examples of such analysis are provided in [Muralidharan et al. \(2016\)](#) and [Muralidharan et al. \(2017b\)](#).

### **5.2.2 Randomization**

Readers are referred to papers dedicated to the subject of randomization procedures and trade-offs associated with them ([Bruhn and McKenzie, 2009](#)) and to the more formal treatment in the companion chapter “The Econometrics of Randomized Experiments” by [Athey and Imbens \(2017\)](#). The main practical point I will add here is that there is a strong case for stratification of randomization (especially by geographic units) to the extent possible. There are several advantages of doing so.

First, it almost always increases power. There is usually considerable spatial heterogeneity in unobservable characteristics (especially when carrying out studies in representative samples as recommended in the previous section). Thus, stratifying randomization at a low-level geographic unit that corresponds to a unit of government administration (such as a district or even a subdistrict) and analyzing the experimental results with stratum fixed effects will soak up a considerable amount of unexplained variation in the outcome variable, and reduce the sum of squared residuals in the regression estimating treatment effects, thereby increasing power (see [Muralidharan and Sundararaman, 2011, 2013, 2015](#); [de Ree et al., 2015](#); [Mbiti et al., 2016](#) for examples). Second, it insures against the risk of other programs being (unexpectedly) implemented in some study areas and not others. While this does not increase bias *ex ante* (since these other programs could be implemented anywhere), it reduces the risk of *ex post* contamination of treatment effects. Since other programs typically get implemented at district or subdistrict levels (especially those by other nongovernmental actors) stratifying and including geographic fixed effects allows the researcher to “net out” these effects. Third, it provides insurance against compliance or data collection problems in a small subset of schools. For instance, it is not uncommon for survey completion rates to be lower in more remote areas, or for treatments to not get implemented in some areas. Instead of having to assume (and justify) that these cases of noncompliance are random, it is often cleaner to just drop the stratum from the analysis (see [de Ree et al., 2015](#) for an illustration).

### **5.2.3 Implementation and follow-up**

It is not uncommon for implementation partners to make changes to the program design or implementation protocols, without realizing that these may compromise the research design (or the interpretation of experimental findings). Ideally, there should be constant contact between researchers and implementation partners to make sure that implementation is on track as intended. If changes are unavoidable then such regular contact can help ensure that (1) changes in implementation protocol do not compromise the evaluation design, and (2) changes in intervention design are clearly documented to enable an accurate description of the program as implemented, which in turn will allow for better

interpretation of the findings. Thus, it is essential to budget and plan for regular monitoring of implementation quality.

### 5.3 Data collection

#### 5.3.1 Outcomes

The main outcomes of education interventions are student participation (enrollment and attendance) and learning outcomes (typically measured by test scores). Attendance is best measured using unannounced visits to schools during the course of the study. If this is not feasible, say for budgetary reasons, another option is to collect attendance data from school records during the time of end-line testing (though these are less reliable). A final option is to use the attendance rate during the end-line tests.

Test scores are the most commonly used outcome measure for RCTs in education, but the standards for psychometric practice in the RCT literature are quite mixed.<sup>27</sup> The default method of measuring treatment effects in education is to express these effects in terms of standard deviations of normalized test scores (normalized relative to the control group). However, the tests used in many education studies are often not designed systematically, and details of test construction are often not reported (even in Appendices). This is problematic because measured treatment effects can be quite sensitive to the sampling of questions from the universe of feasible test questions. For instance, if the test is conducted at a level that is too difficult for most students (floor effects), then measured treatment effects will be zero even if there was a meaningful impact on learning at a level that was below the level at which the test was given (see [Singh, 2015](#) for a discussion of the point, and [Muralidharan et al., 2017b](#) for a demonstration of the importance of this issue in practice). Below, I highlight some important principles of test design that education RCTs should aim to follow:

First, researchers should use tests with a wide distribution of test questions by difficulty. This is often not the case with tests that are designed for rapid assessment of basic learning like the ASER or Uwezo tests (which can have considerable ceiling effects) or grade-appropriate tests (which will typically have large floor effects in developing countries). So it is important that tests be piloted and researchers should ensure that raw test scores are well distributed. Preferably, researchers should select items using Item Response Theory (IRT) and design the most effective tests for the expected ability levels of students. It is not very difficult to do this: testing in three to four schools provides an adequate sample size and involves only a few days of fieldwork. Existing routines in most statistical packages make it easy to generate “Item-Characteristic Curves” (ICCs), and scaled test scores.

<sup>27</sup> In developing countries, the best work on measurement is probably that from the LEAPS studies in Pakistan ([Andrabi et al., 2011](#)) but these are an exception.

A further issue in test design is constructing tests that are comparable across studies. Where possible, tests should have a subset of items which allow them to be linked (through IRT) on a common global distribution of student-learning levels. This will not only enable better comparison across studies (which is essential for cost-effectiveness calculations) but will also allow researchers to place the study sample in the context of the wider distribution of ability in the population. Since most RCTs are not carried out in representative samples, having such information in an Appendix to a paper will be a good practice. Having such linked tests makes it easier to (1) characterize the “business-as-usual” evolution of learning levels in the control group, and (2) express treatment effects in terms of fractions/multiples of the learning in the control group over the same period.<sup>28</sup> Such data also make it possible to characterize heterogeneity in treatment effects in a much richer way than is normally done in education RCTs (see [Muralidharan et al., 2017b](#) for an illustration).

Finally, the main threat to the validity of experimental estimates of the impact of education interventions is the possibility of differential attrition of students between treatment and control groups. High levels of differential attrition can severely compromise the validity of experimental estimates and it is therefore essential to take efforts to minimize the risk of attrition (especially differential attrition across treatment and control groups) during the data collection process (see [Glennerster and Takavarasha, 2013](#) for a detailed discussion).

### **5.3.2 Intermediate inputs, processes, and mechanisms**

The minimum measurement needed for conducting an RCT in education is a set of outcome measures (typically test scores) collected at the end of the study period (assuming that the intervention was successfully randomized). However, the interpretation of experimental results is considerably enriched when accompanied by high-quality data on intermediate inputs, processes, and mechanisms. Opening up the “black box” of treatment effects with such data usually yields much more insight than simply reporting treatment effects ([Kling et al., 2016](#)).

Key intermediate variables to collect data on include school and household expenditure and time use. The importance of the first is illustrated by [Das et al. \(2013\)](#), which was discussed earlier. The importance of the second is illustrated by [Muralidharan and Sundararaman \(2015\)](#) who find in their study of school vouchers that there was no impact on math and native language test scores of winning a voucher and attending a private school. However, they also find that private schools spend much less instructional time

<sup>28</sup> Note that there is one technical challenge in doing such a comparison. The estimated multiple will vary as a function of the test score persistence over time (see [Andrabi et al., 2011](#) and [Muralidharan, 2012](#) for a discussion). One solution to this problem is to present ranges of treatment effects as a function of the persistence parameter (see [de Ree et al., 2015](#) for an application of this approach).

on math and native language and use the time saved to teach other subjects (where they strongly outperform the public schools as may be expected). Thus, the inference on the relative productivity of public and private schools would have been incorrect if the differential patterns of time use had not been accounted for.

More generally, in addition to financial costs, it is also important to consider all opportunity costs of an intervention (which is often not done). Consider the case of modifying curriculum by teaching new content. It is crucial to also specify *what is being replaced* in the existing curriculum to make way for the new additions and to test if there are negative impacts on learning of subjects that may have had their instructional time reduced to make way for the new content. On the other hand, if the new materials are being taught over and above the existing content, it is important to consider the opportunity cost of student and teacher time. This may be low or high, but needs to be accounted for.

A good example of the importance of accounting for time use in schools is provided by [Linden \(2008\)](#) who finds that a computer-enabled instruction program had positive effects on student test scores when offered as an after-school supplementary program, but had negative impacts when it was used to substitute existing teaching activity. Thus, an important lesson for evaluations of education interventions is to account for all costs of the intervention, including *time* and financial costs, and being clear about whether the impact on test scores is coming from additional time on task (either home or school work) or from using existing time more efficiently (either by providing inputs that improve the marginal productivity of time in school, by organizing pedagogy more efficiently, or by reducing slack during the school day). Measuring intermediate inputs to the extent possible is key to enabling such analysis.

Other intermediate variables that can shed light on mechanisms include data on teacher attendance and teaching activity. However, these are difficult to measure precisely within typical research budgets because the former requires multiple unannounced visits for precision, and the latter require extended classroom observation time to meaningfully capture variation in teaching practices and detect changes in teaching practice. However, rich insights into classroom processes can be obtained when such measurement is done well and future research would do well to consider cost-effective tools for measuring classroom practice ([Bruns and Luque 2014; Araujo et al., 2016](#)).

### **5.3.3 Long-term follow-up**

An important limitation of experiments is that their prospective nature makes it difficult to obtain long-term outcomes without waiting for a long period of time. Thus, the majority of education experiments report outcomes within a few years of the program. Nevertheless, it is important for both research and policy to be able to understand the long-term impact of programs, and some of the most influential studies in education and human development have been those that have tracked long-term outcomes including the Perry Preschool study ([Heckman et al., 2010](#)) and the Jamaica home

visitation study ([Gertler et al., 2014](#)). A good example of a more recent experimental intervention with longer term follow-up is provided by [Baird et al. \(2016\)](#) who study the 10-year impacts of the deworming program in Kenya studied by [Miguel and Kremer \(2004\)](#).<sup>29</sup>

In addition to long-term follow-ups of short-duration interventions, a related issue is that of estimating the impacts of treatments that are continued for a long period of time. This is especially relevant for estimating the impacts of changing school-level policies because these changes will affect students for many years (potentially for as many years as they are in school). However, few studies manage to do this for a combination of budgetary and practical reasons. Some exceptions are [Muralidharan and Sundararaman \(2015\)](#) who study the impact of a school choice program after 4 years of exposure to treatment, and [Muralidharan \(2012\)](#) who studies the impact of teacher-performance pay over 5 years of exposure to treatment.

While it is not easy for studies with more limited budgets to plan for either long-term follow-up or long-term experimental exposure to a treatment, it can be extremely valuable to do so. In particular, funders and researchers should try to support long-term follow-ups in cases where the short-term effects are highly encouraging.

## 5.4 Analysis

There are several high-quality existing resources on analysis of experimental data (including [Gerber and Green, 2012](#); and [Athey and Imbens, 2017](#)), and the reader is referred to these ones for a more formal treatment of the topic. This section will briefly highlight issues that are salient for the analysis of experiments in education and outlines a recommended set of analysis for papers in this area to follow, which is typically organized around main treatment effects, heterogeneity, mechanisms of impact, and cost effectiveness.

### 5.4.1 Main treatment effects

A typical estimating equation for studying the impact of receiving an education intervention takes the form as

$$T_{isjk}(Y_n) = \beta_0 + \beta_1 \cdot T_{isjk}(Y_0) + \beta_2 \cdot Treatment_i + \beta_{Z_i} \cdot Z_i + \beta_{X_i} \cdot X_i + \varepsilon_{isjk} \quad (2)$$

where  $T_{isjk}(Y_n)$  represents normalized test scores for student  $i$  in subject  $s$  in grade  $j$  and school  $k$ , at the end of  $n$  years of the experiment. Since test scores are highly correlated

<sup>29</sup> It is worth noting that from a policy perspective, important complementary evidence to [Miguel and Kremer \(2004\)](#) was provided by [Bleakley \(2007\)](#) who presented historical evidence on the long-term positive impacts of a mass deworming program in the US. The combination of short-term experimental evidence and long-term evidence using historical data (albeit less well-identified variation) provided greater confidence in the policy value of launching mass deworming programs.

over time, it is standard to control for baseline test scores to increase the precision of estimates.<sup>30</sup> Including stratum (often geographic)-fixed effects ( $Z_i$ ) helps to absorb geographic variation and increase efficiency, and is needed to account for stratification of the randomization. The main estimate of interest is  $\beta_2$ , which provides an unbiased estimate of the impact of receiving the treatment, and it is standard to estimate  $\beta_2$  both with and without controlling for household socioeconomic characteristics ( $X_i$ ).

Since the treatment effect  $\beta_2$  above is calculated relative to the control distribution (which is a standard normal),  $\beta_0$  will typically be zero (or the omitted fixed effect) and has no cardinal meaning. It is standard to report  $\beta_2$  separately for each subject tested, and to also report the mean treatment effect averaged across all subjects tested to present a summary statistic of impact. Such a summary statistic is especially useful in interpreting studies with positive effects on some subjects and not on others. Since such variation could simply reflect sampling variation (see discussion in the next section on heterogeneity), a summary statistic across subjects is useful to report.<sup>31</sup>

Eq. (2) represents the standard VAM that is the workhorse of the education literature. Note that this VAM does *not* use the intuitive “difference-in-difference” approach (where the dependent variable would be the difference between current and lagged test scores). This is because there is very strong evidence from several settings that test scores are not fully persistent over time. In other words, there is considerable decay in test scores over time and  $\beta_1$  in Eq. (2) is typically estimated to be in the range of 0.3–0.7. The standard “difference-in-difference” specification imposes a restriction that  $\beta_1$  in Eq. (2) equals 1, which is typically rejected in the data. Thus, imposing this restriction would lead to misspecification of the estimating equation and potentially biased estimates of  $\beta_2$ , which is why the default specification in this literature takes the form in Eq. (2). See the excellent discussion of the relevant issues in [Andrabi et al. \(2011\)](#).

Test-score decay (or incomplete persistence) over time is typically not a problem for estimating short-term treatment effects in education experiments, because randomization ensures that mean baseline test scores are comparable across treatment and control schools. However, decay presents challenges when evaluating longer term treatment effects. Specifically, the challenge is that the specification in Eq. (2) can be used to consistently estimate the  $n$ -year effect of the programs (with  $T_{ijk}(Y_0)$  on the right-hand side), but not the “ $n$ ’th” year effect (with  $T_{ijk}(Y_{n-1})$  on the right-hand side) because  $T_{ijk}(Y_{n-1})$

<sup>30</sup> The default baseline score that is controlled for is the score on the same subject and student, but in cases where no baseline test was conducted in the same subject, it is still useful to control for the mean normalized test score across all subjects for which a baseline test was available for the same cohort, or if the cohort did not have a baseline, then the corresponding school-level mean for older cohorts can be included to increase precision (see [de Ree et al., 2015](#) for an illustration).

<sup>31</sup> At the same time, it is also not clear that all subjects should be weighted equally, which is why it is good practice to report results both by subject and averaged across subjects (see [Muralidharan and Sundararaman, 2015](#) for a discussion).

is a posttreatment outcome that will be correlated with the treatment indicator. The literature estimating experimental treatment effects in education therefore typically estimates only the  $n$ -year effect. However, over time, the “loss” of test scores due to decay will typically be higher in treatment schools since they start each year with higher test scores. See [Muralidharan \(2012\)](#) for a discussion of the implications of the distinction between “gross” and “net” treatment effects for the evaluation of education interventions over a longer period of time.<sup>32</sup>

Since the main threat to the validity of an experiment is attrition, the analysis of treatment effects should typically be preceded by a clear description of attrition across treatment and control groups, and a test of equality of attrition levels across treatment and control groups. It is also a good practice to present the student and school-level correlates of attrition and to test whether the same model using observables can predict attrition in both the treatment and control groups (see [Muralidharan and Sundararaman, 2011](#) for an illustration). In cases, where some differential attrition is unavoidable, it is standard to include two kinds of robustness checks. The first is inverse-probability reweighting of observations to recover the distribution of students in the baseline (this is typically only valid if the observable correlates of attrition are similar across treatment and control groups). The second is to use bounding techniques (see [Muralidharan and Sundararaman, 2015](#) for an illustration).<sup>33</sup>

### **5.4.2 Heterogeneity**

Heterogeneous treatment effects are typically estimated with linear interactions across household, school, and teacher covariates, and it is standard to report whether there are differential treatment effects across any of these covariates. Nevertheless, it is important to be cautious in interpreting the results of such analysis for at least two reasons. First, interacting a randomly assigned treatment with a nonrandomly assigned covariate does not provide exogenous variation in the latter.<sup>34</sup> Second, in the absence of a preanalysis plan with well-theorized reasons for heterogeneity along specific dimensions estimated

<sup>32</sup> This discussion is mainly relevant for studies that use test scores as the main dependent variable, and is less relevant for longer term studies that track employment and earnings outcomes.

<sup>33</sup> Of course, the best situation is one that has limited attrition to begin with. This should be a high priority for data collection efforts (see [Glennerster and Takavarashan, 2013](#) for further discussion on how to do so in practice).

<sup>34</sup> A good example is provided in Table 6B of [Muralidharan and Sundararaman \(2011\)](#). They find that teachers with lower base pay responded more to the teacher performance-pay program that they study. These results may suggest that the magnitude of the performance pay mattered because the potential bonus (which was the same for all teachers) from a given level of improvement in student performance would have yielded a larger bonus (as a fraction of base pay) for teachers with lower base pay. However, teacher base pay is also strongly correlated with years of experience, and they find that teachers with fewer years of experience also respond better to the program. This is consistent with the possibility that younger teachers may respond better to any treatment since it may be easier for them to change their behaviors, which is a completely different interpretation of the reason for heterogeneous treatment effects.

heterogeneous effects could simply reflect sampling variation and inference should be corrected for multiple comparisons. One way to make such analysis credible would be to prespecify such heterogeneity in advance (see [Olken, 2015](#) for a discussion).

At the same time, it is also possible that some kinds of treatment heterogeneity that are not anticipated, or prespecified in advance but discovered ex post, may help to make sense of the overall experimental results. Good examples include [Glewwe et al. \(2009\)](#) on the impact of providing free text books in Kenya, and [Muralidharan and Sundararaman \(2015\)](#) on the impact of school choice in India. In the first study, the authors did not specify that they expected stronger effects on students at the top of baseline test score distribution. Nevertheless, finding this result ex post made sense because many of the students with lower baseline test scores were not able to read, which made it unlikely that they would benefit from the provision of a free text book. Similarly, the school-choice experiment studied in [Muralidharan and Sundararaman \(2015\)](#) was not designed to test for heterogeneity by the medium of instruction of the schools that voucher-winning students chose to attend. But the finding (using instrumental variable techniques) that the impact of switching medium of instruction from Telugu (the native language) to English was negative for content subjects was consistent with other research and provided important nuance to understanding the overall experimental results. Thus, it is important to both report such results and to suitably caveat their interpretation.<sup>35</sup>

A particularly useful parameter along which to test for heterogeneity is the baseline student test score, which can be treated as a summary statistic of educational inputs that students had received up to the point when they enter the study. Educational interventions are also well suited to nonparametric analysis of heterogeneity, and doing so as a function of end-line and baseline test score distributions can both be useful ways of characterizing the heterogeneity of program impacts. I describe each approach below.

The first is to estimate quantile treatment effects (defined for each quantile  $\tau$  as:  $\delta(\tau) = G_n^{-1}(\tau) - F_m^{-1}(\tau)$  where  $G_n$  and  $F_m$  represent the empirical distributions of the treatment and control distributions with  $n$  and  $m$  observations, respectively), with bootstrapped confidence intervals for inference. Note that this does *not* plot the treatment effect at different quantiles (since student rank order is not preserved between the baseline and end-line tests even within the same treatment group). It simply plots the gap at each percentile of the treatment and control distributions after the program and compares test scores across treatment and control groups at every percentile of the end-line distribution. Such a plot is especially useful as a visual test of first-order stochastic dominance between treatment and control groups.

<sup>35</sup> As a template for such writing, see [Muralidharan and Sundararaman \(2015\)](#), who start their discussion of heterogeneity of impact by medium of instruction by noting that: “Our experiment was *not* designed to identify heterogeneous effects by school characteristics, but we report some suggestive results that are likely to be important for future research designed explicitly to study such heterogeneity.”

A second way to show heterogeneity is to plot nonparametric treatment effects by percentile of *baseline* score, where the treatment and control end-line distributions are plotted separately with the x-axis being the percentile of baseline score. This plot allows researchers to characterize the treatment effect as a function of where students were in the initial test score distribution (see [Muralidharan and Sundararaman, 2011](#) for a detailed illustration of these two types of analysis). However, this can only be done for cohorts for which baseline data exist.

### **5.4.3 Mechanisms**

Mechanisms of impact are typically shown by comparing data on school and household inputs such as spending, and time-use across treatment and control groups using a similar estimating equation as in Eq. (2). As discussed earlier in this chapter, these can be especially useful for opening up the “black-box” of treatment effects. Illustrative examples of the value of such analysis include: [Muralidharan and Sundararaman \(2011\)](#) for the analysis of changes in teacher behavior in response to a teacher performance pay program; [Das et al. \(2013\)](#) for the analysis of changes in household spending in response to a school grant program; and [Muralidharan and Sundararaman \(2015\)](#) for the analysis of ways in which school-time use differs markedly between public and private schools.

### **5.4.4 Cost effectiveness**

As described in [Section 2.3](#), a unifying theme in the economics of education literature is to compare the relative cost effectiveness of several possible interventions to improve education outcomes. Thus, the final piece of analysis that is recommended is a cost-effectiveness analysis that uses standardized templates for reporting cost (such as recommended by [Dhaliwal et al., 2012](#)) and presents treatment effects in terms of dollars spent per unit test score gain per student. An alternate form of cost-effectiveness analysis that is not done often, but is also very useful is analyzing the effectiveness of interventions per unit of student time spent (see [Muralidharan et al., 2017b](#) for an illustration). While spending on education can (in theory) be augmented continuously, time is finite. Thus, identifying the effectiveness of interventions per unit of time spent is likely to play an important role in improving the productivity of education systems in developing countries.

## **6. CONCLUSION**

The study of education in developing countries has benefited enormously from the rapid growth in field experiments. The most important insight from the experimental research in this area over the past 15 years has been that there is a wide range of cost-effectiveness of education interventions. On the one hand, very expensive policies such as unconditional teacher salary increases have been found to have no impact on learning outcomes.

On the other hand, relatively inexpensive policies like supplemental TaRL with modestly paid and trained volunteers have been found to have large positive impacts on learning outcomes. Overall, the evidence points to several promising ways in which the efficiency of education spending in developing countries can be improved by pivoting public expenditure from less cost-effective categories of expenditure to more cost-effective ways of achieving the same objectives.

At the same time, there are important gaps in our knowledge of how best to design interventions to cost-effectively improve outcomes at scale, and there is much fertile ground for research on education in developing countries. There are important open questions within each of the four categories of interventions summarized in [Section 3](#). On demand, a lot more work is needed on the optimal design of demand-side interventions (beyond showing that they are effective). On inputs, we still do not have good experimental evidence on many important questions including the impacts of improving school infrastructure, teacher-training programs, and class size (the evidence on this is more indirect). On governance, key open questions include understanding the impact of private schools (holding per-student spending constant, and precluding selection of students), and the impact of attempts to improve school governance at scale. In addition to these, I highlight three areas below where the knowledge gaps are large, and where the returns to better evidence are likely to be particularly high.

The first underresearched area where field experiments are likely to yield large returns is pedagogy. Most of the experimental studies on education in developing countries have been conducted by economists, and as a result, the topics on which we have more evidence tend to be topics of interest to economists (such as household demand, information, inputs, and incentives). However, some of the most promising avenues for improving education in low-income settings may involve improving the design and delivery of classroom instruction. While several small-scale innovations may be taking place in this area, there is remarkably little good evidence on the effectiveness of different pedagogical practices in developing countries. For instance, we have very little evidence on how to optimally organize a period of classroom instruction. In other words, the core building block of modern schooling (a period of instruction) is based on a nonexperimental evidentiary standard. This in turn means that the evidence base for the design of teacher training programs is also very limited. While economists may have limited professional incentives to work on domains such as pedagogy, there are likely to be large social returns from researchers trained in designing and running field experiments collaborating with experts in curriculum and pedagogy to improve the empirical evidence-base on effective pedagogy in developing countries.<sup>36</sup>

<sup>36</sup> A good illustration of this point is the consistent evidence on the large gains in student learning obtained from implementing a pedagogical approach that is focused on “Teaching at the Right Level”.

A second underresearched area is post-primary and secondary education. Most of the evidence summarized in this chapter has been from interventions aimed at improving primary education (with the notable exception of conditional cash transfer programs). This is understandable since primary education is foundational and most of the increases in developing country-school enrollment in the past 15 years have been in primary school (consistent with the MDG of achieving universal primary education by 2015). However, the cohorts who benefited from this expansion in primary education are now entering post-primary education in large numbers and there is remarkably little evidence on how to effectively improve the quality of post-primary education. The challenges of effective post-primary education are likely to be considerably greater than those of primary education since the variation in student preparation is likely to be much higher (as shown by [Muralidharan et al., 2017b](#)), and the returns to research here are likely to be high.

A third underresearched area is scale. Specifically, while the evidence summarized in this paper provides a useful starting point in identifying the kinds of interventions that are likely to be effective, we have very little understanding on how to embed these interventions in education systems to deliver improved outcomes at scale. One approach to developing such an understanding is provided by the research program on TaRL led by Abhijit Banerjee and Esther Duflo and conducted in partnership with Rukmini Banerjee of the NGO Pratham for over a decade. Starting with an “efficacy trial” on a small scale (reported in [Banerjee et al., 2007](#)), this research program has featured experimental evaluations of several implementation models in and outside the formal public-school system across several Indian states to better understand how to improve basic literacy and numeracy at scale (see [Banerjee et al., 2010, 2016; Duflo et al. 2015a](#) for a detailed discussion). While the studies reveal several challenges in successfully integrating the successful TaRL pedagogy into the regular education system, the long-term process of iterative program design, implementation, evaluation, and refinement has helped to identify models (such as the learning camps in Uttar Pradesh, and the dedicated TaRL hour in Haryana) that may enable the scaling up of the successful TaRL intervention (see [Banerjee et al., 2016](#) for a detailed discussion).

A second approach is to work with governments to randomize programs during rollout to directly evaluate the impact of education programs at scale. This is the approach I take in ongoing work in the Indian state of Madhya Pradesh where the government agreed to randomize an ambitious program to improve school quality at the scale of 2000 schools. Such an approach may be especially promising when combined with administrative data on outcomes, which sharply reduces the cost of carrying out experiments at scale, and also makes it easier to conduct longer term follow-ups (see [Hastings et al., 2015](#) for an illustration of such a study in Chile).

In addition to the topics mentioned above, researchers would also do well to pay attention to three cross-cutting themes across all categories of education interventions.

These are (1) heterogeneity, (2) data on intermediate variables, and (3) combining experimental and structural methods. A recurring insight in the evidence to date is that optimal policies and interventions are likely to vary as a function of students' initial conditions, and interventions that cater effectively to such heterogeneity are likely to be more effective. A second recurring theme in the discussion in this chapter is the importance of collecting good data on intermediate variables to provide better insights on the mechanisms for program impact (or reasons for lack thereof). A third theme that younger researchers will benefit from paying attention to is the value of embedding experimental interventions in more general models of household behavior. Unifying the "treatment-effects" approach outlined in [Section 2.3](#) and the more structural modeling approach outlined in [Section 2.4](#) is not easy, but if done well, such studies have the potential to expand the research and policy use of experiments, and provide a more informed basis for using experimental results to make predictions regarding the impact of variants of the policy studied.

Field experiments in education in developing countries have been an extremely fertile area of research and policy-relevant insights in the past 15 years. This chapter has aimed to synthesize the most important insights from the existing research and to provide a toolkit for younger researchers embarking on answering the open questions in this area. I expect that the field will continue to be very active, and that it will produce several high-quality studies in the years to come.

## REFERENCES

- Acemoglu, D., 2010. Theory, general equilibrium, and political economy in development economics. *J. Econ. Perspect.* 24 (3), 17–32.
- Altonji, J.G., Mansfield, R.K., 2014. Group-average Observables as Controls for Sorting on Unobservables When Estimating Group Treatment Effects: The Case of School and Neighborhood Effects (NBER Working Paper No. 20781). National Bureau of Economic Research (NBER), Cambridge, MA.
- Andrabi, T., Das, J., Khwaja, A.I., 2015. Report Cards: The Impact of Providing School and Child Test Scores on Educational Markets (Policy Research Working Paper No. 7226). The World Bank, Washington, DC.
- Andrabi, T., Das, J., Khwaja, A.I., Zajonc, T., 2011. Do value-added estimates add value? Accounting for learning dynamics. *Am. Econ. J. Appl. Econ.* 3 (3), 29–54.
- Angelucci, M., De Giorgi, G., 2009. Indirect effects of an aid program: how do cash transfers affect ineligibles' consumption? *Am. Econ. Rev.* 99 (1), 486–508.
- Angrist, J., Bettinger, E., Bloom, E., King, E., Kremer, M., 2002. Vouchers for private schooling in Colombia: evidence from a randomized natural experiment. *Am. Econ. Rev.* 1535–1558. <http://dx.doi.org/10.3386/w8343>.
- Angrist, J., Bettinger, E., Kremer, M., 2006. Long-term educational consequences of secondary school vouchers: evidence from administrative records in Colombia. *Am. Econ. Rev.* 96, 847–862. <http://dx.doi.org/10.1257/aer.96.3.847>.
- Araujo, M.C., Carneiro, P., Cruz-Aguayo, Y., Schady, N., 2016. Teacher quality and learning outcomes in kindergarten. *Q. J. Econ.* 131 (3), 1415–1453.
- Athey, S., Imbens, G.W., 2017. The econometrics of randomized experiments. In: Duflo, E., Banerjee, A. (Eds.), *Handbook of Field Experiments*, vol. 1 (in this volume).

- Atkin, D., 2016. Endogenous Skill Acquisition and Export Manufacturing in Mexico (NBER Working Paper No. 18266). National Bureau of Economic Research (NBER), Cambridge, MA.
- Attanasio, O.P., 2015. The Determinants of Human Capital Formation During the Early Years of Life: Theory, Measurement and Policies. European Economic Association (EEA), Toulouse, France.
- Attanasio, O.P., Meghir, C., Santiago, A., 2012. Education choices in Mexico: using a structural model and a randomized experiment to evaluate Progresa. *Rev. Econ. Stud.* 79 (1), 37–66. <http://dx.doi.org/10.1093/restud/rdr015>.
- Baird, S., Hicks, J.H., Kremer, M., Miguel, E., 2016. Worms at work: long-run impacts of a child health investment. *Q. J. Econ.* 131 (4), 1637–1680.
- Baird, S., McIntosh, C., Ozler, B., 2011. Cash or condition? Evidence from a cash transfer experiment. *Q. J. Econ.* 126, 1709–1753. <http://dx.doi.org/10.1093/qje/qjr032>.
- Baker, G.P., 1992. Incentive contracts and performance measurement. *J. Political Econ.* 598–614.
- Banerjee, A.V., Banerji, R., Berry, J., Duflo, E., Kannan, H., Mukerji, S., Shotland, M., Walton, M., 2016. Mainstreaming an Effective Intervention: Evidence From Randomized Evaluations of “Teaching at the Right Level” in India. Massachusetts Institute of Technology (MIT), Cambridge, MA. Unpublished Manuscript.
- Banerjee, A.V., Banerji, R., Duflo, E., Glennerster, R., Khemani, S., 2010. Pitfalls of participatory programs: evidence from a randomized evaluation in education in India. *Am. Econ. J. Econ. Policy* 2, 1–30. <http://dx.doi.org/10.1257/pol.2.1.1>.
- Banerjee, A.V., Cole, S., Duflo, E., Linden, L., 2007. Remedyng education: evidence from two randomized experiments in India. *Q. J. Econ.* 122, 1235–1264. <http://dx.doi.org/10.1162/qjec.122.3.1235>.
- Barlevy, G., Neal, D., 2012. Pay for percentile. *Am. Econ. Rev.* 102 (5), 1805–1831.
- Barrera-Osorio, F., Bertrand, M., Linden, L.L., Perez-Calle, F., 2011. Improving the design of conditional transfer programs: evidence from a randomized education experiment in Colombia. *Am. Econ. J. Appl. Econ.* 3, 167–195. <http://dx.doi.org/10.1257/app.3.2.167>.
- Barrera-Osorio, F., Linden, L., 2009. The Use and Misuse of Computers in Education: Evidence From a Randomized Experiment in Colombia (Impact Evaluation Series No. 29). The World Bank, Washington, DC.
- Barro, R.J., 1991. Economic growth in a cross section of countries. *Q. J. Econ.* CVI (425), 407–443.
- Beasley, E., Huillery, E., 2012. Empowering Parents in Schools: What They Can(not) Do. Abdul Latif Jameel Poverty Action Lab (J-PAL), Cambridge, MA. Unpublished manuscript.
- Becker, G.S., 1962. Investment in human capital: a theoretical analysis. *J. Political Econ.* 9–49.
- Ben-Porath, Y., 1967. The production of human capital and the life cycle of earnings. *J. Political Econ.* 352–365.
- Benhassine, N., Devoto, F., Duflo, E., Dupas, P., Pouliquen, V., 2013. Turning a shove into a nudge? A “labeled cash transfer” for education. *Am. Econ. J. Econ. Policy* 7, 86–125. <http://dx.doi.org/10.1257/pol.20130225>.
- Berry, J., 2015. Child control in education decisions: an evaluation of targeted incentives to learn in India. *J. Hum. Resour.* 50 (4), 1051–1080.
- Beuermann, D.W., Cristia, J.P., Cruz-Aguayo, Y., Cueto, S., Malamud, O., 2015. Home computers and child outcomes: short-term impacts from a randomized experiment in Peru. *Am. Econ. J. Appl. Econ.* 7 (2), 53–80.
- Bharadwaj, P., Løken, K.V., Neilson, C., 2013. Early life health interventions and academic achievement. *Am. Econ. Rev.* 103 (5), 1862–1891.
- Blattman, C., Niehaus, P., 2014. Show them the money: why giving cash helps alleviate poverty. *Foreign Aff.* 93, 117.
- Bleakley, H., 2007. Disease and development: evidence from hookworm eradication in the American South. *Q. J. Econ.* 122 (1), 73.
- Blimpo, M.P., 2014. Team incentives for education in developing countries: a randomized field experiment in Benin. *Am. Econ. J. Appl. Econ.* 6 (4), 90–109. <http://dx.doi.org/10.1257/app.6.4.90>.
- Bobba, M., Frisancho, V., 2016. Learning About Oneself: The Effects of Signaling Academic Ability on School Choice. Inter-American Development Bank, Washington, DC. Unpublished manuscripts.

- Bobonis, G.J., 2009. Is the allocation of resources within the household efficient? New evidence from a randomized experiment. *J. Political Econ.* 117 (3), 453–503.
- Bobonis, G.J., Finan, F., 2009. Neighborhood peer effects in secondary school enrollment decisions. *Rev. Econ. Stat.* 91 (4), 695–716.
- Bold, T., Kimenyi, M., Mwabu, G., Ng’ang’a, A., Sandefur, J., 2013. Scaling-up what Works: Experimental Evidence on External Validity in Kenyan Education. Center for Global Development, Washington, DC. Unpublished manuscript.
- Borkum, E., He, F., Linden, L.L., 2013. School Libraries and Language Skills in Indian Primary Schools: A Randomized Evaluation of the Akshara Library Program. Abdul Latif Jameel Poverty Action Lab (J-PAL), Cambridge, MA. Unpublished manuscript.
- Bourdon, J., Frölich, M., Michaelowa, K., 2010. Teacher shortages, teacher contracts and their effect on education in Africa. *J. R. Stat. Soc. Ser. A (Stat. Soc.)* 173 (1), 93–116.
- Bruhn, M., McKenzie, D., 2009. In pursuit of balance: randomization in practice in development field experiments. *Am. Econ. J. Appl. Econ.* 1 (4), 200–232.
- Bruno, B., Luque, J., 2014. Great Teachers: How to Raise Student Learning in Latin America and the Caribbean. The World Bank, Washington, DC.
- Burde, D., Linden, L.L., 2013. The effect of village-based schools: evidence from a randomized controlled trial in Afghanistan. *Am. Econ. J. Appl. Econ.* 5, 27–40. <http://dx.doi.org/10.1257/app.5.3.27>.
- Carneiro, P., Heckman, J.J., Vytlacil, E.J., 2011. Estimating marginal returns to education. *Am. Econ. Rev.* 101 (6), 2754–2781.
- Cartwright, N., 2007. Are RCTs the gold standard? *BioSocieties* 2 (1), 11–20.
- Chaudhury, N., Hammer, J., Kremer, M., Muralidharan, K., Rogers, F.H., 2006. Missing in action: teacher and health worker absence in developing countries. *J. Econ. Perspect.* 20 (1), 91–116.
- Chetty, R., Friedman, J.N., Hilger, N., Saez, E., Schanzenbach, D.W., Yagan, D., 2011. How does your kindergarten classroom affect your earnings? Evidence from project STAR. *Q. J. Econ.* 126, 1593–1660. <http://dx.doi.org/10.1093/qje/qjr041>.
- Conn, K., 2014. Identifying Effective Education Interventions in Sub-Saharan Africa: A Meta-analysis of Rigorous Impact Evaluations. Columbia University, New York, NY. Unpublished manuscript.
- Cristia, J., Ibárrarán, P., Cueto, S., Santiago, A., Severín, E., 2012. Technology and Child Development: Evidence From the One Laptop per Child program (Working Paper No. IDB-WP-304). Inter-American Development Bank, Washington, DC.
- Das, J., Dercon, S., Habyarimana, J., Krishnan, P., Muralidharan, K., Sundararaman, V., 2013. School inputs, household substitution, and test scores. *Am. Econ. J. Appl. Econ.* 5, 29–57. <http://dx.doi.org/10.1257/app.5.2.29>.
- de Ree, J., Muralidharan, K., Pradhan, M., Rogers, F.H., 2015. Double for Nothing? Experimental Evidence on the Impact of an Unconditional Teacher Salary Increase on Student Performance in Indonesia. The World Bank, Washington, DC.
- Deaton, A., 2010. Instruments, randomization, and learning about development. *J. Econ. Lit.* 48 (2), 424–455.
- Dehejia, R., Pop-Eleches, C., Samii, C., 2015. From Local to Global: External Validity in a Fertility Natural Experiment. Wagner Graduate School of Public Service, New York, NY.
- Dhaliwal, I., Duflo, E., Glennerster, R., Tulloch, C., 2012. Comparative cost-effectiveness analysis to inform policy in developing countries: a general framework with applications for education. In: Glewwe, P. (Ed.), *Education Policy in Developing Countries*. Abdul Latif Jameel Poverty Action Lab (J-PAL), Cambridge, MA. Unpublished manuscript.
- Dizon-Ross, R., 2016. Parents’ Perceptions and Children’s Education: Experimental Evidence From Malawi. Massachusetts Institute of Technology (MIT), Cambridge, MA. Unpublished manuscript.
- Duflo, E., 2001. Schooling and labor market consequences of school construction in Indonesia: evidence from an unusual policy experiment. *Am. Econ. Rev.* 91, 795–813. <http://dx.doi.org/10.1257/aer.91.4.795>.
- Duflo, E., Berry, J., Mukerji, S., Shotland, M., 2015a. A Wide Angle View of Learning: Evaluation of the CCE and LEP Programmes in Haryana, India (Impact Evaluation Report No. 22). International Initiative for Impact Evaluation (3ie), New Delhi, India.

- Duflo, E., Dupas, P., Kremer, M., 2011. Peer effects, teacher incentives, and the impact of tracking: evidence from a randomized evaluation in Kenya. *Am. Econ. Rev.* 101, 1739–1774. <http://dx.doi.org/10.1257/aer.101.5.1739>.
- Duflo, E., Dupas, P., Kremer, M., 2015b. School governance, teacher incentives, and pupil-teacher ratios: experimental evidence from Kenyan primary schools. *J. Public Econ.* 123, 92–110. <http://dx.doi.org/10.1016/j.jpubeco.2014.11.008>.
- Duflo, E., Glennerster, R., Kremer, M., 2007. Using randomization in development economics research: a toolkit. *Handb. Dev. Econ.* 4, 3895–3962.
- Duflo, E., Hanna, R., Ryan, S.P., 2012. Incentives work: getting teachers to come to school. *Am. Econ. Rev.* 102, 1241–1278. <http://dx.doi.org/10.1257/aer.102.4.1241>.
- Evans, D.K., Popova, A., 2015. What Really Works to Improve Learning in Developing Countries? an Analysis of Divergent Findings in Systematic Reviews (Policy Research Working Paper No. 7203). The World Bank, Washington, DC.
- Fiszbein, A., Schady, N.R., 2009. Conditional Cash Transfers: Reducing Present and Future Poverty. The World Bank, Washington, DC.
- Foster, A.D., Rosenzweig, M.R., 1995. Learning by doing and learning from others: human capital and technical change in agriculture. *J. Political Econ.* 1176–1209.
- Ganimian, A.J., Murnane, R.J., 2016. Improving educational outcomes in developing countries: lessons from rigorous evaluations. *Rev. Educ. Res.* XX (X), 1–37.
- Gerber, A.S., Green, D.P., 2012. Field Experiments: Design, Analysis, and Interpretation. W.W. Norton, New York, NY.
- Gertler, P., Heckman, J., Pinto, R., Zanolini, A., Vermeesch, C., Walker, S., Grantham-McGregor, S., 2014. Labor market returns to an early childhood stimulation intervention in Jamaica. *Science* 344 (6187), 998–1001.
- Glennerster, R., Takavarsha, K., 2013. Running Randomized Evaluations: A Practical Guide. Princeton University Press.
- Glewwe, P., Hanushek, E.A., Humpage, S.D., Ravina, R., 2014. School resources and educational outcomes in developing countries: a review of the literature from 1990 to 2010. In: Glewwe, P. (Ed.), *Education Policy in Developing Countries*. University of Chicago Press, Chicago, IL and London, UK.
- Glewwe, P., Ilias, N., Kremer, M., 2010. Teacher incentives. *Am. Econ. J. Appl. Econ.* 2, 205–227. <http://dx.doi.org/10.1257/app.2.3.205>.
- Glewwe, P., Kremer, M., Moulin, S., 2009. Many children left behind? Textbooks and test scores in Kenya. *Am. Econ. J. Appl. Econ.* 1, 112–135. <http://dx.doi.org/10.1257/app.1.1.112>.
- Glewwe, P., Kremer, M., Moulin, S., Zitzewitz, E., 2004. Retrospective vs. prospective analyses of school inputs: the case of flip charts in Kenya. *J. Dev. Econ.* 74, 251–268. <http://dx.doi.org/10.1016/j.jdeveco.2003.12.010>.
- Glewwe, P., Maiga, E.W., 2011. The impacts of school management reforms in Madagascar: do the impacts vary by teacher type? *J. Dev. Eff.* 3 (4), 435–469. <http://dx.doi.org/10.1080/19439342.2011.604729>.
- Glewwe, P., Muralidharan, K., 2016. Improving school education outcomes in developing countries: evidence, knowledge gaps, and policy implications. In: *Handbook of Economics of Education*.
- Green, J.R., Stokey, N.L., 1983. A comparison of tournaments and contracts. *J. Political Econ.* 349–364.
- Hastings, J., Neilson, C.A., Zimmerman, S.D., 2015. The Effect of Earnings Disclosure on College Enrollment Decisions (Working Paper No. 21300). National Bureau of Economic Research (NBER), Cambridge, MA.
- Heckman, J.J., Moon, S.H., Pinto, R., Savelyev, P.A., Yavitz, A., 2010. The rate of return to the HighScope Perry Preschool program. *J. Public Econ.* 94 (1), 114–128.
- Heckman, J.J., Smith, J.A., 1995. Assessing the case for social experiments. *J. Econ. Perspect.* 9 (2), 85–110.
- Hirshleifer, S., 2015. Incentives for Effort or Outputs? a Field Experiment to Improve Student Performance. University of California at San Diego, San Diego, CA. Unpublished manuscript.
- Holmstrom, B., Milgrom, P., 1987. Aggregation and linearity in the provision of intertemporal incentives. *Econ. J. Econ. Soc.* 303–328.
- Holmstrom, B., Milgrom, P., 1991. Multitask principal-agent analyses: incentive contracts, asset ownership, and job design. *J. Law Econ. Organ.* 7, 24–52.
- Itoh, H., 1991. Incentives to help in multi-agent situations. *Econ. J. Econ. Soc.* 611–636.

- Jensen, R., 2010. The (perceived) returns to education and the demand for schooling. *Q. J. Econ.* 125, 515–548. <http://dx.doi.org/10.1162/qjec.2010.125.2.515>.
- Jensen, R., 2012. Do labor market opportunities affect young women's work and family decisions? Experimental evidence from India. *Q. J. Econ.* 127, 753–792. <http://dx.doi.org/10.1093/qje/qjs002>.
- Kandel, E., Lazear, E.P., 1992. Peer pressure and partnerships. *J. Political Econ.* 801–817.
- Kandori, M., 1992. Social norms and community enforcement. *Rev. Econ. Stud.* 59 (1), 63–80.
- Kling, J., Ludwig, J., Congdon, B., Mullainathan, S., 2016. Social policy: mechanism experiments and policy evaluations. In: Duflo, E., Banerjee, A. (Eds.), *Handbook of Field Experiments*. North Holland.
- Kremer, M., 1993. The O-ring theory of economic development. *Q. J. Econ.* 551–575.
- Kremer, M., 2003. Randomized evaluations of educational programs in developing countries: some lessons. *Am. Econ. Rev.* 93 (2), 102–106.
- Kremer, M., Brannen, C., Glennerster, R., 2013. The challenge of education and learning in the developing world. *Science* 340, 297–300. <http://dx.doi.org/10.1126/science.1235350>.
- Kremer, M., Miguel, E., Thornton, R., 2009. Incentives to learn. *Rev. Econ. Stat.* 91, 437–456. <http://dx.doi.org/10.1162/rest.91.3.437>.
- Kremer, M., Sarychev, A., 2008. Why Do Governments Operate Schools? Harvard University, Cambridge, MA. Unpublished manuscript.
- Krishnaratne, S., White, H., Carpenter, E., 2013. Quality Education for All Children? What Works in Education in Developing Countries (Working Paper No. 20). International Initiative for Impact Evaluation (3ie), New Delhi, India.
- Lai, F., Luo, R., Zhang, L., Huang, X., Rozelle, S., 2015. Does computer-assisted learning improve learning outcomes? Evidence from a randomized experiment in migrant schools in Beijing. *Econ. Educ. Rev.* 47, 34–48. <http://dx.doi.org/10.1016/j.econedurev.2015.03.005>.
- Lakshminarayana, R., Eble, A., Bhakta, P., Frost, C., Boone, P., Elbourne, D., Mann, V., 2013. The Support to Rural India's Public Education System (STRIPES) trial: a cluster randomised controlled trial of supplementary teaching, learning material and material support. *PLoS One* 8 (7), e65775.
- Lalive, R., Cattaneo, M.A., 2009. Social interactions and schooling decisions. *Rev. Econ. Stat.* 91 (3), 457–477.
- LaLonde, R.J., 1986. Evaluating the econometric evaluations of training programs with experimental data. *Am. Econ. Rev.* 604–620.
- Lassibille, G., Tan, J.-P., Jesse, C., Nguyen, T.V., 2010. Managing for results in primary education in Madagascar: evaluating the impact of selected workflow interventions. *World Bank Econ. Rev.* 1–27. <http://dx.doi.org/10.1093/wber/lhq009>.
- Lazear, E.P., 2006. Speeding, terrorism, and teaching to the test. *Q. J. Econ.* 1029–1061.
- Lazear, E.P., Rosen, S., 1981. Rank-order tournaments as optimum labor contracts. *J. Political Econ.* 89 (5), 841–864.
- Linden, L.L., 2008. Complement or Substitute? The Effect of Technology on Student Achievement in India. Abdul Latif Jameel Poverty Action Lab (J-PAL), Cambridge, MA. Unpublished manuscript.
- Loyalka, P., Liu, C., Song, Y., Yi, H., Huang, X., Wei, J., Rozelle, S., 2013. Can information and counseling help students from poor rural areas go to high school? Evidence from China. *J. Comp. Econ.* 41, 1012–1025. <http://dx.doi.org/10.1016/j.jce.2013.06.004>.
- Lucas, R.E., 1988. On the mechanics of economic development. *J. Monet. Econ.* 22 (1), 3–42.
- Lucas, R.E., 1990. Why doesn't capital flow from rich to poor countries? *Am. Econ. Rev.* 80 (2), 92–96.
- Majumdar, T., 1983. *Investment in Education and Social Choice*. Cambridge University Press.
- Malamud, O., Pop-Eleches, C., 2011. Home computer use and the development of human capital. *Q. J. Econ.* 126, 987–1027. <http://dx.doi.org/10.1093/qje/qjr008>.
- Mankiw, N.G., 2006. The macroeconomist as scientist and engineer. *J. Econ. Perspect.* 20 (4), 29–46.
- Mankiw, N.G., Romer, D., Weil, D.N., 1992. A contribution to the empirics of economic growth. *Q. J. Econ.* 107 (2), 407–437.
- Mbiti, I., Muralidharan, K., Romero, M., Schipper, Y., Manda, C., Rajani, R., 2016. Inputs, Incentives, and Complementarities in Primary Education: Experimental Evidence From Tanzania. University of California at San Diego, San Diego, CA. Unpublished manuscript.

- McEwan, P., 2014. Improving learning in primary schools of developing countries: a meta-analysis of randomized experiments. *Rev. Educ. Res.* XX, 1–42. <http://dx.doi.org/10.3102/0034654314553127>.
- Miguel, E., Kremer, M., 2004. Worms: identifying impacts on education and health in the presence of treatment externalities. *Econometrica* 72, 159–217. <http://dx.doi.org/10.1111/j.1468-0262.2004.00481.x>.
- Moretti, E., 2004. Workers' education, spillovers, and productivity: evidence from plant-level production functions. *Am. Econ. Rev.* 94 (3), 656–690.
- Muralidharan, K., 2012. Long-term Effects of Teacher Performance Pay: Experimental Evidence From India. University of California, San Diego, San Diego, CA. Unpublished manuscript.
- Muralidharan, K., 2013. Priorities for primary education policy in India's 12th five-year plan. *India Policy Forum* 2012–13 9, 1–46.
- Muralidharan, K., 2016. A new approach to public sector hiring in India for improved service delivery. *India Policy Forum* 12, 187–225.
- Muralidharan, K., Das, J., Holla, A., Mohpal, A., 2017a. The fiscal cost of weak governance: evidence from teacher absence in India. *J. Public Econ.* 145, 116–135.
- Muralidharan, K., Niehaus, P., Sukhtankar, S., 2016. Building state capacity: evidence from biometric smart-cards in India. *Am. Econ. Rev.* 106 (10), 2895–2929.
- Muralidharan, K., Prakash, N., 2016. Cycling to school: increasing secondary school enrollment for girls in India. *Am. Econ. J. Appl. Econ.* (forthcoming). University of California at San Diego, San Diego, CA.
- Muralidharan, K., Singh, A., Ganimian, A.J., 2017b. Disrupting Education? Experimental Evidence on Technology-Aided Instruction in India, 2017 (NBER Working Paper 22923). National Bureau of Economic Research (NBER), Cambridge, MA.
- Muralidharan, K., Sundararaman, V., 2010. The impact of diagnostic feedback to teachers on student learning: experimental evidence from India. *Econ. J.* 120, F187–F203. <http://dx.doi.org/10.1111/j.1468-0297.2010.02373.x>.
- Muralidharan, K., Sundararaman, V., 2011. Teacher performance pay: experimental evidence from India. *J. Political Econ.* 119, 39–77. <http://dx.doi.org/10.1086/659655>.
- Muralidharan, K., Sundararaman, V., 2013. Contract Teachers: Experimental Evidence From India (Working Paper No. 19440). National Bureau of Economic Research (NBER), Cambridge, MA.
- Muralidharan, K., Sundararaman, V., 2015. The aggregate effect of school choice: evidence from a two-stage experiment in India. *Q. J. Econ.* 130 (3), 1011–1066. <http://dx.doi.org/10.1093/qje/qjv013>.
- Muralidharan, K., Zieleniak, Y., 2014. Chasing the Syllabus: Measuring Learning Trajectories in Developing Countries With Longitudinal Data and Item Response Theory. University of California, San Diego, San Diego, CA. Unpublished manuscript.
- Neal, D., Schanzenbach, D.W., 2010. Left behind by design: proficiency counts and test-based accountability. *Rev. Econ. Stat.* 92 (2), 263–283.
- Olken, B.A., 2015. Promises and perils of pre-analysis plans. *J. Econ. Perspect.* 29 (3), 61–80.
- Pradhan, M., Suryadarma, D., Beatty, A., Wong, M., Gaduh, A., Alisjahbana, A., Artha, R.P., 2014. Improving educational quality through enhancing community participation: results from a randomized field experiment in Indonesia. *Am. Econ. J. Appl. Econ.* 6, 105–126. <http://dx.doi.org/10.1257/app.6.2.105>.
- Pritchett, L., Sandefur, J., 2013. Context Matters for Size: Why External Validity Claims and Development Practice Don't Mix (Working Paper No. 336). Center for Global Development (CGD), Washington, DC.
- Reinikka, R., Svensson, J., 2004. Local capture: evidence from a central government transfer program in Uganda. *Q. J. Econ.* 119 (2), 679–705.
- Sabarwal, S., Evans, D., Marshak, A., 2014. The Permanent Input Hypothesis: The Case of Textbooks and (No) Student Learning in Sierra Leone (Policy Research Working Paper No. 7021). The World Bank, Washington, DC.
- Sen, A., 1993. Capability and well being. In: Sen, A., Nussbaum, M. (Eds.), *The Quality of Life*. Oxford University Press, Oxford, United Kingdom.
- Singh, A., 2015. Private school effects in urban and rural India: panel estimates at primary and secondary school ages. *J. Dev. Econ.* 113, 16–32.

- Snistveit, B., Stevenson, J., Menon, R., Phillips, D., Gallagher, E., Geleen, M., Jobse, H., Schmidt, T., Jimenez, E., 2016. The impact of education programmes on learning and school participation in low- and middle-income countries: a systematic review summary report. In: 3ie Systematic Review Summary 7. International Initiative for Impact Evaluation (3ie), London.
- Todd, P.E., Wolpin, K.I., 2003. On the specification and estimation of the production function for cognitive achievement. *Econ. J.* 113 (485), F3–F33.
- Todd, P.E., Wolpin, K.I., 2006. Assessing the impact of a school subsidy program in Mexico: using a social experiment to validate a dynamic behavioral model of child schooling and fertility. *Am. Econ. Rev.* 96 (5), 1384–1417.
- Yang, Y., Zhang, L., Zeng, J., Pang, X., Lai, F., Rozelle, S., 2013. Computers and the academic performance of elementary school-aged girls in China's poor communities. *Comput. Educ.* 60 (1), 335–346. <http://dx.doi.org/10.1016/j.compedu.2012.08.011>.

## CHAPTER 4

# Social Policy: Mechanism Experiments and Policy Evaluations<sup>a</sup>

W.J. Congdon<sup>\*1</sup>, J.R. Kling<sup>§¶</sup>, J. Ludwig<sup>¶||</sup>, S. Mullainathan<sup>¶\*\*</sup>

\*ideas42, New York, NY, United States

§Congressional Budget Office, Washington, DC, United States

¶NBER (National Bureau of Economic Research), Cambridge, MA, United States

||University of Chicago, Chicago, IL, United States

\*\*Harvard University, Cambridge, MA, United States

<sup>1</sup>Corresponding author: E-mail: bill@ideas42.org

## Contents

1. Introduction	390
2. What Are Mechanism Experiments?	394
2.1 Definition	394
2.2 Some history	396
3. Why Do Mechanism Experiments?	397
3.1 Concentrating resources on the parameters with the most uncertainty	399
3.2 Ruling out policies (and policy evaluations)	401
3.3 Complement other sources of evidence	403
3.4 Understand the role of context in moderating policy effects	405
3.5 Expand the set of policies for which we can forecast effects	407
4. When to Do Mechanism Experiments Versus Policy Evaluations?	409
4.1 Mechanism experiment is sufficient	411
4.1.1 A single mechanism	412
4.1.2 Multiple (but not too many) candidate mechanisms	415
4.2 Mechanism experiment plus policy evaluation	416
4.2.1 Mechanism experiments then policy evaluation	417
4.2.2 Policy evaluation followed by mechanism experiment	419
4.3 Just do policy evaluation	420
5. Conclusion	422
References	423

<sup>a</sup> This chapter was prepared for the *Handbook on Field Experiments*, edited by Abhijit Banerjee and Esther Duflo, and draws heavily from: Jens Ludwig, Jeffrey R. Kling and Sendhil Mullainathan, 2011, “Mechanism experiments and policy evaluations,” *Journal of Economic Perspectives*, 25 (3): 17–38. For excellent research assistance we thank Laura Brinkman and Michael Reddy. We thank Nava Ashraf, David Autor, Iwan Barankay, Jon Baron, Howard Bloom, Lorenzo Casaburi, Philip Cook, Stefano DellaVigna, John DiNardo, Esther Duflo, Judy Gueron, Elbert Huang, Chad Jones, Lawrence Katz, Supreet Kaur, John List, Stephan Meier, David Moore, Steve Pischke, Harold Pollack, Dina Pomeranz, David Reiley, Frank Schilbach, Robert Solow, Timothy Taylor, and conference participants at the University of Pennsylvania’s Wharton School of Business, the American Economic Association, and the NBER Handbook of Field Experiments Conference for helpful comments. For financial support, we thank the Russell Sage Foundation (through a visiting scholar award to Ludwig). Any errors and all opinions are our own. The views expressed here are those of the authors and should not be interpreted as those of the Congressional Budget Office.

## Abstract

Policymakers and researchers are increasingly interested in using experimental methods to inform the design of social policy. The most common approach, at least in developed countries, is to carry out large-scale randomized trials of the policies of interest, or what we call here *policy evaluations*. In this chapter, we argue that in some circumstances the best way to generate information about the policy of interest may be to test an intervention that is different from the policy being considered, but which can shed light on one or more key mechanisms through which that policy may operate. What we call *mechanism experiments* can help address the key external validity challenge that confronts all policy-oriented work in two ways. First, mechanism experiments sometimes generate more policy-relevant information per dollar of research funding than can policy evaluations, which in turn makes it more feasible to test how interventions work in different contexts. Second, mechanism experiments can also help improve our ability to forecast effects by learning more about the way in which local context moderates policy effects, or expand the set of policies for which we can forecast effects. We discuss how mechanism experiments and policy evaluations can complement one another, and provide examples from a range of social policy areas including health insurance, education, labor market policy, savings and retirement, housing, criminal justice, redistribution, and tax policy. Examples focus on the US context.

## Keywords

Field experiment; Program evaluation; Randomized controlled trial

## JEL Code

C90

## 1. INTRODUCTION

Randomized experiments have a long tradition of being used in the United States to test social policy interventions in the field, dating back to the social experimentation that began in the 1960s.<sup>1</sup> The use of field experiments to test social policies has accelerated in recent years. For example the US Department of Education in 2002 founded the Institute for Education Sciences with a primary focus on running experiments, with an annual budget that was \$574 million in 2015 ([US Department of Education, 2015](#)). This trend has been spurred in part by numerous independent groups that promote policy experimentation.<sup>2</sup>

This trend toward ever-greater use of randomized field experiments has led to a vigorous debate within economics about the value of experimental methods for informing policy (e.g., [Angrist and Pischke, 2009, 2010](#); [Banerjee and Duflo, 2009](#);

<sup>1</sup> [Gueron and Rolston \(2013\)](#), along with the chapter in this volume by Gueron, provide an account of this early period in the development of randomized demonstration projects for social policy.

<sup>2</sup> Examples include the Campbell Collaboration, the Jameel Poverty Action Lab at MIT, the University of Chicago Urban Labs, the Lab for Economic Opportunity at Notre Dame University, and the Laura and John Arnold Foundation.

(Deaton, 2010; Heckman, 2010; Imbens, 2010). There is little disagreement that a well-executed experimental test of a given policy carried out in a given context provides a strong claim to internal validity—differences in outcomes reflect the effects of the policy within the experimental sample itself. The debate instead focuses on concerns about external validity—that is, to what other settings can the result of a given field experiment be generalized.

In the area of social policy and in many other areas, this debate has often been framed as a choice between experimental and nonexperimental methods. But this ignores an important choice of how to employ experimental methods that we argue here deserves greater attention. Specifically, in this chapter we argue (and demonstrate through numerous examples) that—perhaps counter-intuitively—the best way to test a policy is not always to directly test the policy of interest. Greater use could be made of randomized field experiments that test mechanisms of action through which social policies are hypothesized to affect outcomes—what we call *mechanism experiments*—even if the interventions tested do not directly correspond to those policies we are interested in understanding.

An example may help to illustrate our argument. Suppose the US Department of Justice (DOJ) wanted to help local police chiefs decide whether to implement “broken windows” policing, which is based on the theory that police should pay more attention to enforcing minor crimes like graffiti or vandalism because they can serve as a “signal that no one cares,” and thereby accelerate more serious forms of criminal behavior (Kelling and Wilson, 1982, p. 31). Suppose that there is no obviously exogenous source of variation in the implementation or intensity of broken windows policing across areas, which rules out the opportunity for a study of an existing “natural experiment” (Meyer, 1995; Angrist and Pischke, 2009). To an experimentally minded research economist, the most obvious next step would be for DOJ to choose a representative sample of cities, randomly assign half to receive broken windows policing, and carry out what we would call a traditional *policy evaluation*.

Now consider an alternative experiment: Buy a small fleet of used cars. Break the windows of half of them. Randomly select neighborhoods and park the cars there, and measure whether more serious crimes increase in response. While this might initially seem like a fanciful idea, this is basically the design that was used in a 1960s’ study by the Stanford psychologist Philip Zimbardo (as described by Kelling and Wilson, 1982, p. 31). The same idea was used more recently by Keizer et al. (2008) who examined the effects of various forms of disorder (such as graffiti or illegal firecrackers exploding) and found substantially more litter and theft occurred when they created disorder. One can of course perform variants with other small crimes, or randomly select neighborhoods for the reduction of disorder such as clean-up of smashed liquor bottles, trash, and graffiti. This *mechanism experiment* does not test a policy: it directly tests the causal mechanism that underlies the broken windows policy.

Which type of experiment would be more useful for public policy? The underlying issue is partly one of staging. Suppose the mechanism experiment failed to find the causal mechanism operative. Would we even need to run a policy evaluation? If (and this is the key assumption) we are confident that a policy implementing broken windows policing would affect crime only by reducing disorder and were convinced that we had strong evidence that reducing disorder does not affect crime, then we could stop. Running the far cheaper mechanism experiment first serves as a valuable screen. Conversely, if the mechanism experiment found strong effects, we might run a policy evaluation to figure out how much disorder could be reduced by applying broken windows policing at different levels of intensity. Indeed, depending on the costs of the policy evaluation, the magnitudes found in the mechanism experiment, and what else we think we already know about the policing and crime “production functions,” we may even choose to adopt the policy straightaway.

In our example, mechanism experiments help us stretch research funding further, which bears directly on the ability of experimentation to create generalizable knowledge that is useful for social policy. One way to address external validity with randomized field experiments is replication—that is, testing the policy in many different contexts. As [Angrist and Pischke \(2010, pp. 23, 24\)](#) argue, “a constructive response to the specificity of a given research design is to look for more evidence, so that a more general picture begins to emerge … the process of accumulating empirical evidence is rarely sexy in the unfolding, but accumulation is the necessary road along which results become more general.”<sup>3</sup> One challenge to this strategy stems from resource constraints. In the broken windows application, mechanism experiments help with these resource constraints by incorporating prior knowledge and letting us focus on the issues about which the most remains to be learned.

In the spirit of contributing to a handbook that is intended to be of practical use to both policymakers and researchers, we organize the remainder of this chapter into three sets of applied questions: In [Section 2](#) we clarify and expand on the answer to the question: What are mechanism experiments? Mechanism experiments can be defined broadly as tests of the causal mechanisms ( $M$ ) that link policies ( $P$ ) to social outcomes ( $Y$ ). Mechanism experiments test the  $M \rightarrow Y$  relationship using interventions that do not necessarily correspond to the actual policies of immediate interest. The connection of a mechanism to a clearly specified policy, not just to social outcomes, helps distinguish what we mean by mechanism experiments from more general efforts within economics to understand what determines outcomes.

In [Section 3](#) we answer the question: Why do mechanism experiments? A primary motivation, as noted above, is to help establish external validity. Mechanism experiments can do this in two ways. First, they can increase the amount of policy-relevant information

<sup>3</sup> As [Cook and Campbell \(1979\)](#) note, “tests of the extent to which one can generalize across various kinds of persons, settings and times are, in essence, tests of statistical interactions … In the last analysis, external validity … is a matter of replication” (p. 73, 78).

per research dollar available, since replication of policy evaluations is a costly way to learn about external validity. Mechanism experiments can concentrate resources on parameters where policymakers have the most uncertainty, as in the broken windows example, or help us rule out policy evaluations that we don't need to run, or for that matter rule out policies, with the added benefit of sometimes reducing the amount of time required to realize that some policy is not actually promising. Second, mechanism experiments address questions of external validity related to forecasting the contexts in which a policy would have effects. Mechanism experiments can improve our ability to forecast effects by learning more about the way in which local context moderates policy effects, or expand the set of policies for which we can forecast effects.

In Section 4, we answer the questions: When should we do a mechanism experiment, when should we do a policy evaluation, and when should we do both? One necessary condition for doing a mechanism experiment is that researchers or policymakers need to believe they know at least something about the candidate mechanisms through which a policy affects social welfare. If the list of candidate mechanisms is short and the costs of carrying out a full-blown policy evaluation are high (or if the policy stakes are low), a mechanism experiment by itself might be sufficient to inform policy. Likely to be more common are situations in which it makes sense to follow a mechanism experiment with a policy evaluation to understand other links in the causal chain from policy to outcomes, or to calibrate magnitudes. The mechanism experiment still adds great value in these cases by helping us prioritize resources for those areas where a full-blown policy evaluation is worth doing. We note that in some situations, such as when there is a long list of candidate mechanisms that could have interactive effects or even work at cross-purposes, it may not be worth doing a mechanism experiment and researchers should just proceed to carry out a black-box policy evaluation.

While our discussion largely focuses on the key conceptual points behind our argument, we also try to illustrate the potential contributions (and limitations) of mechanism experiments with existing social policy studies whenever possible. As we discuss further below, at present mechanism experiments are relatively more common in developing than developed country contexts, partly because for a variety of reasons development experiments are more typically carried out with NGOs than with government partners. The potential gains from rebalancing the policy experiment portfolio to include more mechanism experiments, not just policy evaluations, seem largest within the developed country context. Partly for that reason, we focus most of our discussion and examples on the developed country context with which we are most familiar ourselves, the United States. For a more comprehensive summary of social policy experiments that have been carried out in the United States, see [Greenberg and Shroder \(2004\)](#).<sup>4</sup>

<sup>4</sup> An updated version of their publication *The Digest of Social Experiments* is in progress.

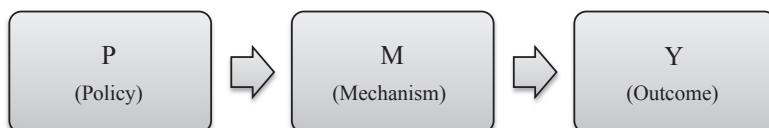
## 2. WHAT ARE MECHANISM EXPERIMENTS?

### 2.1 Definition

Broadly, a mechanism experiment is an experiment that tests a mechanism—that is, it tests not the effects of variation in policy parameters themselves, directly, but the effects of variation in an intermediate link in the causal chain that connects (or is hypothesized to connect) a policy to an outcome. That is, where there is a specified policy that has candidate mechanisms that affect an outcome of policy concern, the mechanism experiment tests one or more of those mechanisms. There can be one or more mechanisms that link the policy to the outcome, which could operate in parallel (for example when there are multiple potential mediating channels through which a policy could change outcomes) or sequentially (if for example some mechanisms affect take-up or implementation fidelity). The central idea is that the mechanism experiment is intended to be informative about some policy but does not involve a test of that policy directly.

In our broken windows example, given above, one could sketch this model as follows: Policing policies ( $P$ ) that target and reduce minor offenses such as broken windows ( $M$ ) ultimately lead to reductions in the thing policymakers are most concerned about serious criminal offenses ( $Y$ ) (Fig. 1). The hypothetical policy evaluation in that case—randomly assign cities to receive broken windows policing, and track outcomes for more serious crimes—is a policy evaluation, a test of  $P \rightarrow Y$ . The result tells policymakers whether that policy has an impact on the outcomes of key policy interest. The corresponding mechanism experiment—randomly assign cars with broken windows across neighborhoods—is a test of  $M \rightarrow Y$ . It tells policymakers whether the mechanism is operative.

Even though the mechanism experiment does not resemble in any way the policy of interest, it can concentrate resources on estimating the parameters most relevant to policy decisions, leading the experiment to be informative for policy to a surprising degree. Suppose that from previous work policymakers also know the elasticity of minor offenses with respect to policing ( $P \rightarrow M$ ), and they also believe that change in minor offenses is the only mechanism through which broken windows may affect the outcome of ultimate policy concern (serious offenses). What policymakers do not know is the accelerator: by how much will reducing minor offenses cascade into reducing serious offenses. The mechanism experiment estimates the parameter about



**Figure 1** Policies, mechanisms, and outcomes.

which there is the greatest uncertainty or disagreement ( $M \rightarrow Y$ ). In contrast, the policy evaluation that measures the policy's impact on serious crimes,  $P \rightarrow Y$ , also provides information about the crime accelerator, but with more noise because it combines the variability in crime outcomes with the variability in the impact of policing on minor crimes in any given combination of cities and years. With enough sample (that is, money) one could recover the  $M \rightarrow Y$  link.

The broken windows example is not an isolated case: many policies have theories built into them, even if they are sometimes left implicit. Often these theories can be tested more cost effectively and precisely with experiments that do not mimic the policy of direct interest, or in some cases do not even mimic anything likely to ever be a real (or even feasible) policy. But while social scientists already value mechanism experiments because they contribute to building knowledge, we argue that *even if the sole goal were informing policy* (social policies or others), mechanism experiments, even those that do not test real policy options, play a crucial and underappreciated role.

In our broken windows example, it is of course possible that a policymaker who is initially interested in the question of whether broken windows policing works sees the results of a mechanism experiment and decides that a policy to (say) remediate blight in distressed neighborhoods might itself be worth supporting at scale. The mechanism experiment in this case has the effect of reorienting policymaker attention to expand the set of policies ( $P$ 's) they consider. But the key definitional point is: What we would consider a mechanism experiment depends on the initial policy question. Our goal is to help policymakers, policy-research funders and policy-oriented researchers see that if the focus is on broken windows policing, the experiment that yields the most useful information per dollar spent relevant to broken windows policing does not necessarily involve the delivery of broken windows policing.

Note that this definition does not imply that mechanism experiments are necessarily “cheap” in an absolute sense, only in a relative sense compared to other policy evaluations that would produce the same amount of policy-relevant information. For example, the RAND health insurance experiment was a mechanism experiment by our definition, since it was intended to tell us something about how cost-sharing provisions of health insurance plans affect health but included a treatment arm that did not correspond to a real (or even feasible) policy—a condition with zero co-insurance ([Newhouse and the Insurance Experiment Group, 1993](#)). Yet as we will argue below, this “extreme dosage” arm makes it possible for policymakers to interpolate the effects of plans with a wide range of co-insurance rates at cost lower than would be required to carry out at-scale policy evaluations of plans at each possible co-insurance rate. At a total cost of \$285 million in 2010 dollars, the RAND experiment also holds the record—for now—as the most expensive mechanism experiment ever ([Greenberg and Shroder, 2004](#), p. 181).

## 2.2 Some history

The use of mechanism experiments is not new in the US context, even if the label has been developed only relatively recently (Ludwig et al., 2011b). In fact, some of the earliest social policy experiments in the United States took the form of what we would classify as mechanism. For example, the New Jersey Income Maintenance Experiment is often considered the first large-scale social policy experiment in the United States (Greenberg and Shroder, 2004). This experiment was explicitly concerned more with testing mechanisms (behavioral responses to different marginal tax rates, including negative tax rates) than with the study of existing policies. This experiment and the RAND health insurance experiment mentioned previously were not isolated examples (even if they were not labeled as mechanism experiments at the time they were carried out).

In fact, a critique some economists leveled against the large-scale government social experiments of the 1970 and 1980s was precisely that they were “not necessarily test[ing] real policy options” (Harris, 1985, p. 161). In addition, budget changes in the 1980s reduced support for policy-relevant social experimentation of any sort. For these and other reasons, mechanism experiments of this sort fell out of favor in the United States over the ensuing decades.

Mechanism experiments continue to be relatively common at present in the developing world context, which has seen a substantial growth in policy-oriented experimentation over the past decade (Banerjee and Duflo, 2009). Of course there are large-scale policy evaluations that occur in developing country contexts, such as the *Progresa* conditional cash transfer experiment in Mexico (see for example Skoufias et al., 2001; Rivera et al., 2004; Schultz, 2004). There are also investigator-initiated policy evaluations, such as tests of ordeal mechanisms for receipt of social program assistance (Alatas et al., 2013). But the ratio of mechanism experiments to policy evaluations is higher in developing than developed countries.

One candidate explanation is that public sector capacity is typically more limited in developing countries, and so development experiments tend to be carried out frequently in partnership with NGOs rather than with government agencies. For example, Ashraf et al. (2010) study the role of prices in mediating the use of social services by sending privately hired marketers out to households to offer them a water chlorination product at different offered prices that are then transacted at a second randomized price. The policy of interest is the price at which the product is sold to households, which can affect usage and hence health ( $Y$ ) through two candidate mechanisms—a screening mechanism ( $M_1$ ) that affects which households use the product at different prices, and a sunk-cost mechanism ( $M_2$ ) that suggests families may be more willing to use the product if they have paid more for it. The double randomization of the initial offer price and ultimate transaction price leads to an intervention that does not exactly correspond to an actual

policy, but which does help separate out these two mechanisms and so help shed light on the optimal pricing and subsidy policies (in the same spirit also see for example [Karlan and Zinman, 2009](#); [Cohen and Dupas, 2010](#)).

Our focus on the rest of the chapter will be on the developed world context, particularly the developed country we know best (the United States), since such a large share of the randomized experiments carried out in the United States that are intended to be helpful to social policy involve directly testing actual policies, often carried out in collaboration with government agencies. To take just one recent example, the experimental evaluation of the Reemployment and Eligibility Assessment (REA) initiative funded by the US Department of Labor was a large-scale test of a policy already in place in most states ([Poe-Yamagata et al., 2011](#)). The US Department of Education's Institute for Education Sciences (IES) describes its goals in funding experiments to be “development of practical solutions for education from the earliest design stages through pilot studies and rigorous testing at scale,” and “large-scale evaluations of federal education programs and policies.” Numerous private research firms around the country are carrying out multiple multimillion dollar policy evaluations at any point in time with support from private foundations and government. The resources and attention devoted today in the United States and other developed nations to what we could consider mechanism experiments pales in comparison to what is devoted to policy evaluations. In the next section, we develop our argument for why we believe this balance should shift over time.

### **3. WHY DO MECHANISM EXPERIMENTS?**

In a given setting, which policy P generates the greatest change in outcomes Y at a given cost? That is the central question of primary interest to policymakers. Given that objective, why carry out mechanism experiments, which do not test actual (or perhaps even feasible) policies?

The answers are motivated partly by the inevitable question we have with any policy evaluation, which has to do with its external validity. The effects of, say, broken windows policing in Chicago's affluent North Shore suburb of Evanston may differ from what would happen as a result of this intervention in the distressed neighborhoods of Chicago's south side. We are worried that the “treatments” we study may interact with attributes of the target population, context, or time period. These baseline attributes that interact with treatment effects are what non-economists call “moderators.”

This concern that treatments may interact with context has led naturally to the view that the best way to produce generalizable information is to focus on policy interventions of the sort that policymakers might actually implement, and test them in multiple settings of the sort in which the policy might actually be implemented. One way to think about what we are trying to accomplish through this replication comes from the useful distinction suggested by [Wolpin \(2007\)](#) and [Todd and Wolpin \(2008\)](#) between *ex-post policy*

*evaluation*—understanding what happened as the result of a policy or program that was actually implemented—and *ex-ante policy evaluation*, which, as DiNardo and Lee (2011, p. 2) describe it, “begins with an explicit understanding that the program that was actually run may not be the one that corresponds to a particular policy of interest. Here, the goal is not descriptive, but instead predictive. What would be the impact if we expanded eligibility of the program? What would the effects of a similar program be if it were run at a national (as opposed to a local) level? Or what if the program were run today (as opposed to 20 years ago)? It is essentially a problem of forecasting or extrapolating, with the goal of achieving a high degree of external validity.”

Replicating tests of real policies in different contexts tells us something about the value of using the policy’s average effect as a forecast for what we would expect to accomplish in other settings. An obvious challenge with this approach is that policy evaluations are expensive and often difficult to carry out. One use of mechanism experiments is to increase the policy-relevant information we obtain for a given research budget, to maximize the coverage of policy-relevant contexts about which we have some information. Mechanism experiments help us do this by:

- Concentrating resources on parameters where there is the most uncertainty,
- Ruling out policies (and the need for full-blown policy evaluations), and
- Extracting more information from other (nonexperimental) sources of evidence.

Of course, evidence of treatment effect heterogeneity is not fatal to the idea of using policy experiments to help inform policy, since it is always possible to use forecasting methods that emphasize results from settings that are similar to whatever local context is being considered for some new policy. For example, we might predict the effects of broken windows policing in south side Chicago by focusing on results from Evanston’s poorer neighborhoods specifically. Forecasting becomes essentially a matching or reweighting exercise (see, for example, Hotz et al., 2005; Cole and Stuart, 2010; Imbens, 2010; Stuart et al., 2011). The value of replicating tests of real policies comes from the fact that the chances of finding a “match” for any future policy application increases with the number of policy-relevant contexts in which the actual policy has been tested. Mechanism experiments can generate useful information for this type of policy forecasting exercise in two main ways:

- Understanding mechanisms of action can help shed light on those contextual factors that moderate policy effects, and so help forecast policy effects to different contexts, and
- Expanding the set of policies for which we can forecast policy effects by testing extreme policy “doses,” so that forecasting relies on interpolation not extrapolation.

In the remainder of this section, we discuss these different uses for mechanism experiments in greater detail and include several examples. Because mechanism experiments remain underutilized in developed-country contexts, we present several hypothetical examples that illustrate the potential value-added of this approach for the study of social

problems in places like the United States. Where possible we also present real examples of what we consider to be mechanism experiments, even if the authors themselves might not have explicitly set out to execute a mechanism experiment when they launched their studies.

### **3.1 Concentrating resources on the parameters with the most uncertainty**

Social policymaking invariably involves operating along a causal chain of variable length. Policy reforms are reflected in statutory or regulatory changes, leading to corresponding adjustments in program administration and implementation, to which individuals or other actors respond along sometimes multiple margins. The result of what happens along the full causal chain is what ultimately ends up determining social welfare impacts. At each step, the impacts are uncertain, especially to the extent that ultimate impacts depend on behavioral responses.

Of course the option of testing the entire chain jointly through a full policy evaluation is always available. But depending on what we already know about some of the links in the chain, this might not be the most *efficient* way to learn about the likely effects of a policy. Suppose there is a policy ( $P$ ) that is thought to affect some outcome ( $Y$ ) through a candidate mediator or mechanism ( $M$ ), so that the theory of change behind the intervention is expressed as:  $P \rightarrow M \rightarrow Y$ . If we already believe we understand the  $P \rightarrow M$  link, for instance, we can concentrate resources on understanding the remaining part of the theory of change for the policy without having to incur the costs of a full policy experiment.

In this way, we can use field experiments to learn about social policy design without necessarily testing actual policies. A mechanism experiment will allow us to identify the response of  $Y$  to  $M$ , and, in combination with what we already believed we knew about  $P \rightarrow M$ , this allows us to learn what we ultimately want to know—about  $P \rightarrow Y$ —without having to test the full policy or every point in the logic chain. Under the most straightforward conditions—there is only a single candidate  $M$ , and the relationship between  $M \rightarrow Y$  is stable—this boils down to, essentially: we can learn about the sign of the response of  $Y$  to  $M$ , the magnitude of the response of  $Y$  to  $M$ , and the shape of the response of  $Y$  to  $M$ .

For example, there is a great deal of concern right now in the United States about the quality of police-community relations, and in particular with the prevalence of use of force by police officers against civilians (particularly in high-crime, economically and racially segregated communities). One hypothesis for this use of force is officer exhaustion—many police officers wind up, working second jobs to make ends meet—which, combined with long and sometimes unpredictable hours at their main job, can lead to overtired officers whose ability to cope with stressful situations is impaired. A potential policy response would be to substantially increase police salaries. Suppose

we were confident that big pay raises for police would reduce the number of second jobs they worked and reduce exhaustion (that is, we understood the  $P \rightarrow M$  link) but we were unsure of the degree to which less-exhausted officers would help improve police-community interactions and reduce police use of force against civilians ( $M \rightarrow Y$ ).<sup>5</sup> Since union rules in most cities discourage horizontal inequities in treatment of officers, it would be hard to carry out a policy evaluation that randomized officers for pay raises within cities. Instead, any policy evaluation of pay raises for police may well need to be carried out with the police department as the unit of random assignment, which would be an enormously costly way to learn about the overall effects of the policy ( $M \rightarrow Y$ ).

Now consider an alternative mechanism experiment: In any big city at any point in time there are officers coming back from a week or two of vacation (“furlough”). Randomly select some police beats but not others within the city to be assigned an officer just returning from furlough, then see what happens to measures of police use of force (or citizen complaints against the police) in those beats that were randomly assigned officers who should be better-rested. This experiment does not test the policy of interest (increased pay), but does tell us something about the causal link we are most unsure of—between the hypothesized mechanism and the outcome of interest—at greatly reduced cost compared to a full-blown policy evaluation. Indeed by demonstrating that manipulation of this candidate mechanism can have causal impacts on the outcome of interest, this mechanism experiment may suggest other policies that operate through that mechanism (reduced officer exhaustion) as candidate policies.

Mechanism experiments can go beyond answering questions about the sign or size of the  $M \rightarrow Y$  link and also illuminate the shape of that relationship. For example, one candidate mechanism through which stepped-up police resources may reduce crime is an increase in the likelihood that victims report to the police (as suggested by Levitt, 1998). However, suppose we did not know whether the effect of additional victim reporting (the  $M \rightarrow Y$  link) gets larger or smaller as the overall level of victim reporting changes; that is, we did not know whether the effect of victim reporting on crime is convex or concave. A very costly way to test this hypothesis is to carry out full-blown policy evaluations that assign increased police presence to some neighborhoods but not others across a large number of neighborhoods. A lower-cost way to test this hypothesis would be a mechanism experiment that randomly assigned rewards for victim reports to the police that result in an arrest in some areas but not in other areas. By exploiting either naturally occurring variation across areas in baseline victim reporting rates, or by

<sup>5</sup> If we were unsure about the effects of the policy (higher salaries) on officers working second jobs, for the sake of our example one could imagine a policy that offered the police union higher salaries in exchange for an agreement that officers would not take on second jobs (the same way that many professional athletes have written into their contracts that they cannot engage in other activities like skiing that can endanger their health).

randomly varying the size of the rewards across areas, we could learn about the functional form for the relationship between  $M \rightarrow Y$ . That would then help us prioritize where to carry out our policy evaluations—that is, where we expect the effect of police on crime to be largest.

### 3.2 Ruling out policies (and policy evaluations)

Mechanism experiments can also lead to efficiency gains where they can obviate the need for policy experimentation or development by ruling out candidate policies without requiring direct tests of full implementation of those policies. Consider the case where a policy  $P$  is under consideration because of a theoretical link to an outcome of interest  $Y$ , which is hypothesized to be mediated by mechanism  $M$  (or, as in the notation above,  $P \rightarrow M \rightarrow Y$ ). Where the uncertainty is around  $M \rightarrow Y$ , rather than inferring the relationship from analysis of  $P \rightarrow Y$ , we can just test  $M \rightarrow Y$  directly. If  $M$  is not linked to  $Y$ —and assuming away for the moment any other candidate mechanism by which  $P$  could affect  $Y$ —we can rule out policies that we hypothesize operate through that  $M$ , and move on to more promising avenues of inquiry.

Consider, for example, a policy design question that comes up from time to time related to the Earned Income Tax Credit (EITC), which is whether the EITC would better support goals associated with poverty alleviation and work promotion if it were structured as an earnings supplement, rather than as, in practice, an earnings-linked lump sum transfer. This question is often raised in the context of a larger policy debate around, more generally, whether income supports like the EITC would be better structured as wage subsidies (Phelps, 1994).

Consider a policy,  $P$ , for example, to restructure the terms and delivery of the EITC to mimic more closely a wage subsidy, with the goal of improving the welfare,  $Y$ , of beneficiaries by advancing the payments and promoting consumption smoothing in a way that we think will increase their utility, dollar for dollar. A policy evaluation that randomly enrolled individuals into either the current EITC or this new policy could answer questions about the welfare impacts of this policy, but would be administratively difficult, expensive, and require overcoming difficult measurement challenges. A mechanism experiment on this same issue sheds important light on the question: researchers encouraged take-up of a little used (and no longer available) option for workers to receive their EITC in earlier and more frequent payments known as the Advance EITC. Promoting this option to beneficiaries, imposing deadline on the choice, and requiring active choice of the way of receiving the credit did not actually increase take-up of the Advance EITC in this sample (Jones, 2010). Taking that revealed preference as an indication that taking up the Advance EITC would not have increased welfare, the experiment provides evidence for policymakers that

smoothing payments of the EITC, or consumption out of the EITC, may not be a worthwhile policy change.<sup>6</sup> (In fact, partly based on evidence like this, the Advance EITC was dismantled in 2011.) Recipients do not appear to want to use the EITC in this way; rather, beneficiaries seem to prefer to make use of the lump sum nature of the credit as a form of forced saving.

In addition to ruling out policies, mechanism experiments can also in some cases possibly obviate the need for full-blown policy evaluations. The example from the introduction about an experiment testing the mechanism behind the broken windows theory might provide evidence that would be a basis for forgoing an evaluation of a policy intervention aimed at reducing minor offenses, depending upon the results.

In the case of one of the large-scale social policy experiments from a few decades ago, which studied the national Job Training Partnership Act (JTPA), a federally implemented policy for promoting employment and earnings among dislocated adults and economically disadvantaged adults and youth was evaluated nationally (Bloom et al., 1997). The full program evaluation randomly assigned 21,000 eligible individuals to either receive JTPA services, or not. Under the policy, P, the services provided by JTPA varied by local provider, but generally focused on skill development, and included classroom training, on-the-job training, and other forms of job training. The mechanisms by which the policy was supposed to operate were varied, but very much centered on the idea that the skills and credentials conferred by this type of training—as typified by the general education diploma (GED), receipt of which was in many cases the focus of the training—would allow beneficiaries to command a higher wage. The outcomes of interest were employment and earnings. The evaluation found no positive impact for youth, and only modest positive benefits for adults.

Although it is impossible to know for certain, based on what we now know from other research, this seems like a case where a well-designed mechanism experiment could potentially have at least called into question the need for the full evaluation of a policy such as this. Work by Heckman et al. (2011) finds that the credential of the GED and the type of skills that it reflects are not well correlated with the skills that command wage premia in labor markets, including those paying lower wages. A mechanism experiment could potentially have been designed that examined whether the skills provided through JTPA were valued in the labor market—say in a study where resumes with the sorts of degrees, test scores, and descriptions of skills that would be fostered by JTPA training were sent to employers. If resumes appearing to have JTPA training did not generate greater interest among employers than other resumes, this would have been a signal that a JTPA-style policy evaluation may have been unnecessary.

<sup>6</sup> A new, ongoing experiment with periodic EITC payments (Bellisle and Marzahl, 2015) reaffirms at least the administrative feasibility of an advance EITC.

### 3.3 Complement other sources of evidence

Experimental evidence has many desirable properties for informing policy, but it is necessarily part of a larger portfolio of policy-relevant evidence. Field experiments exist in the context of other important and useful sources of evidence for informing social policy, including not just policy evaluations but also nonexperimental sources of evidence such as natural or quasi-experiments. Randomized experiments and natural experiments may be complements in a broader program of research on an issue that involves multiple stages (Kling, 2007).

We have argued that one important part of the value of mechanism experiments is to help increase the amount of policy-relevant information that can be obtained for a given research budget, by testing interventions that might not look like an actual (or even feasible) policy. One way mechanism experiments can do that is by increasing the amount of information we can extract from other types of policy-oriented research. Mechanism experiments can help us interpret the results of policy evaluations and quasi-experimental work, including null findings. Once we know that some mechanism is linked to an outcome, the first thing we would check upon seeing a zero impact in a full-scale policy evaluation is whether the policy successfully changed the mediator. Evidence that the mediator was unchanged would suggest the potential value of testing other policies that might generate larger changes in the mediator.

To take an example, consider the conflicting and largely inconclusive body of evidence related to the performance and achievement impacts of school-choice policies (Rouse, 1998; Hoxby, 2003; Cullen et al., 2006; Figlio and Rouse, 2006). Much of this evidence is from quasi-experimental work, although some of it is experimental. The economic theory for the mechanism by which greater choice of schools should lead to improved academic outcomes is fairly straightforward: given greater choices, parents (or whoever is making schooling decisions) can optimize over a choice set of schools, with respect to academic outcomes, and schools can respond to the competitive pressures that are generated. Overall, while the result should be that greater choice leads to improved academic outcomes, the evidence of such effects is scattered. There are a number of points along the causal chain at which this logic could fail to hold, but reduced-form evidence on the effectiveness of school choice does not identify which particular mechanism(s) do not operate as hypothesized.

One mechanism experiment that helps shed light on this mixed evidence was performed by Hastings and Weinstein (2008), who provided actionable, simplified information on school quality to parents. Given that information, parents in the treatment group tended to choose schools with higher test scores. This result unpacks, and provides evidence for, a potential mechanism bottleneck that could explain weak results from other sources of evidence on the effects of school choice. Parents might not have the information necessary, or be able to parse available information effectively, in order to select

better performing schools for their children. Moreover, if parents are not doing this effectively, then schools may not be responding to parental choices, either.

Or consider another case, where a policy,  $P$ , is intended to affect a particular mechanism,  $M$ , under the theory that it mattered for  $Y$ . A null finding from evidence looking at the effects of  $P$  on  $Y$  might occur because  $P$  failed to actually affect  $M$ , but also might be because  $M$  is not linked to  $Y$ . A mechanism experiment that shows whether  $M$  does or does not matter for  $Y$  resolves this.

Even in the case where policy evaluation or quasi-experimental evidence finds that a policy is successful in achieving an outcome of interest, complementary mechanism experiments might still be informative for policy design. New mechanism experiments could also be designed with the explicit goal of better understanding existing natural experiment findings.

For instance, suppose policymakers are concerned that high levels of violence impair the ability of children growing up in distressed urban neighborhoods to succeed in school. This hypothesis is suggested by a series of clever nonexperimental studies that reanalyze population surveys that administer achievement tests to study subjects, and take advantage of the fact that respondents (who hail from different neighborhoods) are assessed at different points in time. Being assessed shortly after a violent event (such as a homicide) occurs in one's neighborhood substantially reduces achievement test scores—on the order of 0.5–0.66 standard deviations in one study (Sharkey, 2010). The size of these effects is enormous, given that for example the black-white test score gap nationwide is typically estimated to be on the order of 0.4–0.8 standard deviations depending on the subject area and age at which tests are administered.<sup>7</sup>

The findings suggest that any policy ( $P$ ) that reduces violence should improve at least short-term academic outcomes ( $Y$ ). Unfortunately, this quasi-experimental design is not well suited for telling us about the outcome of primary policy concern—long-term academic outcomes. A policy evaluation that tried to answer this question could become quite costly given the need to administer an intervention substantial enough to reduce crime in distressed neighborhoods and keep it in place for the long term. Such an evaluation would also need some way to deal with the complication of how to measure long-term changes in exposure to violence given residential mobility in and out of the target neighborhoods.

Now consider the following mechanism experiment: One plausible mechanism ( $M$ ) for the link between exposure to violence and academic outcomes is the effect of crime on stress (Buka et al., 2001). Imagine we identified a sample of people living in

<sup>7</sup> For example the black-white test score gap among 13-year olds in the United States in math is about 0.8 standard deviations in the National Assessment of Educational Progress. On the other hand, the gap measured in the Early Childhood Longitudinal Study of Kindergarteners in reading skills is about 0.4 standard deviations when children start school (Fryer and Levitt, 2004).

high-crime neighborhoods and randomly assigned some to receive long-term enrollment in a meditation-based stress-reduction program (Kabat-Zinn et al., 1992) and then tracked how children did in school over time.<sup>8</sup>

### 3.4 Understand the role of context in moderating policy effects

A central question facing social policy experimenters is the issue of when and how to export results across contexts. This type of policy forecasting, in which the effects of a policy are estimated before it is put in place, will inevitably require more assumptions, theory, and guesswork than studies on policies that have already been tried (see also [Harrison and List, 2004](#), p. 1033). But policy forecasting is in the end at least as important for public policy. As the distinguished physicist [Richard Feynman \(1964\)](#) once argued, “The moment you make statements about a region of experience that you haven’t directly seen, then you must be uncertain. But we always must make statements about the regions that we have not seen, or the whole business is no use.” Put differently, in order to forecast the effects of a policy for a new population or in some new geographic context or time period, we need to understand something about the policy’s moderators, which can sometimes be facilitated by mechanism experiments that identify mechanisms of actions.

On a practical level, mechanism experiments present a less costly and more practical way to generate direct empirical evidence about the stability of interventions across contexts. Mechanism experiments can be lower-cost ways of understanding how the  $P \rightarrow Y$  link varies across contexts by letting us focus resources on understanding how  $M \rightarrow Y$  link varies across contexts when the  $M \rightarrow Y$  link is the most uncertain link in the causal chain.

Consider for example the US Department of Housing and Urban Development’s (HUD’s) Moving to Opportunity (MTO) residential-mobility experiment. Since 1994, MTO has enrolled around 4600 low-income public housing families with children and randomly assigned them into three groups: (1) a *traditional voucher group*, which received a standard housing voucher that subsidizes them to live in private-market housing; (2) a *low-poverty voucher group* that received a standard housing voucher that is similar to what was received by the traditional voucher group, with the exception that the voucher could only be redeemed in Census tracts with 1990 poverty rates

<sup>8</sup> Mechanism experiments can also help us build on natural experiment studies by better understanding how to improve policies. That is, if we have a policy that has lots of candidate M’s, we could use mechanism experiments to isolate the relative importance of these to design new policies in future that focus more on (and up the dosage for) the key M’s. In the previous example, suppose we were unsure about whether exposure to violence mattered because of stress or instead because of, say, depression. We could complement the natural experiment study with two mechanism experiments: one focused on stress (such as meditation) and the other on addressing depression (for example by providing pharmacotherapy). We discuss the possibility of multiple mechanisms in more detail below.

below 10%; and (3) a *control group*, which received no additional services. Assignment to the low-poverty voucher group led to more sizable changes in neighborhood poverty and other neighborhood characteristics than did assignment to the traditional voucher group (Sanbonmatsu et al., 2011). While the traditional voucher arm has the form of a policy experiment, the low-poverty arm had the form of a mechanism experiment—it involved a location restriction that is not a realistic policy option in the US context.

MTO was found to have important effects on both physical and mental health (Kessler et al., 2014; Ludwig et al., 2011a, 2012, 2013; Sanbonmatsu et al., 2011).<sup>9</sup> But the experimental evidence in MTO leaves the precise mechanism generating those effects unidentified, so it is easy to speculate that the causal pathways include mechanisms that either are or are not likely to demonstrate high external validity. So, if those effects happened to operate through, say, something about differences between urban and suburban policing in 1990s in the selected set of cities, we might think the external validity of those results may not be high. If, however, we were able to isolate precisely that MTO effects were due to reductions in experienced stress, that alone improves our ability to make an out of sample forecast because we then more precisely know that what we have to consider is how invariant the relationship is between physical or mental health and stress.

Many “behaviourally informed” policy interventions that are motivated by the view that people are often imperfect decision-makers can be viewed as mechanism experiments that help to elucidate the role of context in policy outcomes. Consider for example the policy question of how the EITC changes individual income, that is, whether the EITC is an effective form of redistribution. One link in the causal chain involves the decision to claim the EITC. Because of the way the policy is implemented in this instance, as a tax credit, the outcome—the degree of income support it provides to recipients—is mediated by claiming and receipt of the credit. Indeed, we observe that there are eligible individuals who fail to receive the credit (and so, any corresponding benefits)—even among individuals who already file income taxes, for whom the marginal costs of claiming the credit are incidental, a portion do not claim the credit (Plueger, 2009). We also observe considerable variation across local areas in how individuals respond to the EITC (Chetty et al., 2013).

One hypothesis about why take-up varies has to do with the potentially moderating effects of variation in information about the credit and knowledge of eligibility on individual use of and response to the credit. In part, because this particular method of income support is administered through the tax code, however, conducting a full-blown policy evaluation experimenting with different versions of the EITC would be impractical for

<sup>9</sup> The same pattern generally holds in the follow-up of MTO outcomes measured 4–7 years after baseline; see Kling et al. (2005), Sanbonmatsu et al. (2006), Kling et al. (2007), and Fortson and Sanbonmatsu (2010).

the purposes about learning about the role of information. But from a mechanism experiment sending timely, simplified notices we see that such an intervention can lead to increased claiming ([Bhargava and Manoli, 2013](#); [Manoli and Turner, 2014](#)). This experiment used reminders to test the impact of changes to the mechanism, that is, claiming and receipt of the credit. Simple mailings to a set of roughly 35,000 individuals who appeared eligible for the EITC but did not claim the credit led to significant increases in receipt. However, the effects of these notices faded rapidly and dramatically. In this way, we learn about the extent to which the policy outcomes of the EITC are mediated by claiming frictions generated from the way the credit is currently administered. Policymakers can draw conclusions based on this result for the design of the EITC—most directly, for the way in which the terms of the credit are communicated to eligible individuals, but also potentially for the information requirements those terms reflect.

### **3.5 Expand the set of policies for which we can forecast effects**

Mechanism experiments are not constrained in the same way that policy evaluations are to testing actually feasible or implementable versions of social policies so they can, as a result, test extreme parameter values or unusual functional forms of interventions. Testing unrealistically intensive or pure treatment arms has the benefit of letting us forecast the effects of a wide range of more realistic policy options in those cases when our policy experiments do identify successful interventions. As [Hausman and Wise \(1985, pp. 194–95\)](#) noted 30 years ago: “If, for policy purposes, it is desirable to estimate the effects of possible programs not described by treatments, then interpolations can be made between estimated treatment effects. If the experimental treatments are at the bounds of possible programs, then of course this calculation is easier.”

As a result, while these types of experimentation in social policy can sometimes be viewed as uninformative or irrelevant to policy design, the opposite is the case: by generating information on the nature and range of the behavioral response to an aspect of a policy, mechanism experiments can expand the set of policies for which we can accurately forecast effects. Mechanism experiments can provide a low-cost way to deliver large, even extreme, doses of M to see if the  $M \rightarrow Y$  link matters, and to get a sense of responsiveness of Y to M. By way of comparison, if policy evaluations are constrained to implementable variants of P, and so only manipulate M within the restricted range that allows given the  $P \rightarrow M$  relationship, our understanding of how the policy did or did not work may be inconclusive. If our experiments test interventions that are as intensive as (or even more intensive than) anything that could be accomplished by actual policies, and still don’t work, this lets us rule out policies, as well.

The policy impact that this type of study can have is illustrated by the RAND Health Insurance Experiment that was introduced above ([Newhouse and the Insurance](#)

[Experiment Group, 1993](#)). Run from 1971 to 1982, this experiment randomly assigned 2750 families to different styles and levels of health insurance coverage. The experiment was designed to provide information on the social welfare impacts of health insurance coverage. The intermediate outcome of interest was behavioral response to health insurance—visits to doctors, hospitals, etc.—and the ultimate outcomes of interest were health outcomes themselves. The central findings were that utilization of health care was responsive to cost sharing, and that overall cost sharing did not have strong effects on health outcomes (though there were some negative effects for lower income participants).

Most notably, the RAND experiment included many treatment arms that do not correspond to any sort of health insurance policy one could buy today. The most generous treatment arm in the RAND experiment offered essentially free coverage, with zero percent coinsurance; other arms were 25%, 50%, and 95% coinsurance rates. Yet this now-decades-old experiment remains one of our most important sources of information about how the generosity of health insurance plans affects the demand for health care and subsequent health outcomes.<sup>10</sup> It continues to be cited heavily even in modern health insurance policy debates, and instrumental to the experiment's prolonged usefulness is the fact that, as a mechanism experiment, it was able to generate such substantial variation in cost-sharing terms in order to observe and estimate behavioral responses.

As another example, in MTO assignment to the low-poverty voucher group led to more sizable changes in neighborhood poverty and other neighborhood characteristics than did assignment to the traditional voucher group ([Ludwig et al., 2008](#)). Aside from a few important physical and mental health outcomes, overall, the traditional voucher treatment had relatively few impacts on MTO parents or children through 10–15 years after baseline ([Ludwig et al., 2011a, 2012, 2013](#); [Sanbonmatsu et al., 2011](#)). While the low-poverty voucher treatment did not have the sweeping impacts across all outcomes that would be predicted by much of the sociological literature, low-poverty vouchers did generate substantial changes in adult mental and physical health outcomes and overall well-being, had mixed effects on a number of youth outcomes—with girls doing generally better on a number of measures while boys did not. For children who moved to lower poverty neighborhoods when they were relatively young, the treatment led to long-run positive impacts on earnings ([Chetty et al., 2016](#)).

Three of us (Congdon, Kling, and Ludwig) have worked on MTO for many years, and have often heard the reaction that the traditional voucher treatment is more policy-relevant and interesting than the low-poverty voucher treatment, because only the

<sup>10</sup> While it was of modest size, it was not cheap. At a total cost of \$285 million in 2010, the RAND experiment also holds the record—for now—as the most expensive mechanism experiment ever ([Greenberg and Shroder, 2004](#), p. 181).

former corresponds to a realistic policy option. But it was the low-poverty voucher that generated a sufficiently large “treatment dose” to enable researchers to learn that *something* about neighborhood environments *can* matter for many of these important outcomes, a fact that would not have been discovered if MTO’s design had only included the more realistic traditional voucher treatment. The finding from the low-poverty voucher also provides lessons for why the standard voucher policy has not had such effects (at least in part, it appears, by not inducing sufficient mobility at least in socioeconomic terms). For this reason, findings from the low poverty voucher arm of the experiment have been very influential in housing policy circles.

To take a final example, a policy option sometimes considered to protect workers against a loss of earning power late in their careers is wage-loss insurance (Davidson, 1995; Kletzer and Litan, 2001; LaLonde, 2007). Under most designs of wage-loss insurance, the policy replaces, for covered workers who have lost their job and find reemployment only at a lower wage, some portion of the difference between their older and new wage. The optimal way to set the replacement rate parameter is a question of direct interest for policymakers and researchers. That rate should be set to balance goals of promoting reemployment and supporting consumption while not discouraging search or human capital development. But how individuals will respond is ultimately an empirical question.

In many proposals, the replacement rate is set at 50%; this was also the rate set in a wage insurance demonstration implemented under the *Trade Adjustment Assistance* program. One of the most useful pieces of evidence for informing the design of this policy, however, has been the results of a Canadian experiment with wage-loss insurance that set a replacement rate of 75% (Bloom et al., 1999). That experiment found that covered workers returned to work somewhat faster, but possibly at lower wages. This mechanism experiment testing the functional form of the policy under consideration but with parameter values not under consideration was able to provide information about response patterns that is still useful for policymakers. It is relatively straightforward to interpret the implications of that result for a policy with a 50% replacement, by interpreting the finding as an elasticity. But using relatively extreme values of the policy parameter made it more likely the experiment would precisely estimate a point on the response curve. Even at the larger value of 75%, responses were modest; if a policy experiment had tested a lower replacement rate, the (presumably) relatively smaller responses to the replacement rate would have been harder to detect.

#### **4. WHEN TO DO MECHANISM EXPERIMENTS VERSUS POLICY EVALUATIONS?**

The purpose of our paper is not to argue that economists should do only mechanism experiments, or that mechanism experiments are in any sense better than policy evaluations.

Our point instead is that given the relative paucity of mechanism experiments, there may be value in having economists do more of them.

**Table 1** presents a framework for thinking about when mechanism experiments can help inform policy decisions. In order for a mechanism experiment to make any sense, we need to believe that we know something about the candidate mechanisms through which a policy might affect outcomes (the key contrast across the columns of **Table 1**). For a mechanism experiment to be able to tell us something useful about a policy, or to be able to help inform investment of research funding across different candidate policy

**Table 1** Policy experiment checklist

	Prior beliefs/understanding of mechanisms	
	Low	High
Implications for experimental design	<p>Run a policy evaluation. (or)</p> <p>Do more basic science; use multiple methods to uncover mechanisms.</p>	<p>Run a mechanism experiment to rule out policies (and policy evaluations). (or)</p> <p>Run mechanism experiment to help rule in policies. Either follow with full policy evaluation (depending on costs of policy evaluation, and potential program benefits/scale), or use results of mechanism experiment for calibration and structural estimation for key parameters for benefit-cost calculations.</p>
Implications for policy forecasting/external validity	<p>Run multiple policy evaluations; carry out policy forecasting by matching to estimates derived from similar policies and settings (candidate moderators).</p> <p>Debate: Which characteristics to match on? where do these come from?</p>	<p>Use mechanism knowledge to measure characteristics of policy and setting (moderators) for policy forecasting.</p> <p>Can run new mechanism experiments to test in different settings prior to carrying out policy evaluations in those settings.</p>

evaluations, we either need the list of candidate mechanisms to be “not too long” or to believe that the candidate mechanisms will not interact or work at cross-purposes. Otherwise, information about the causes or consequences of just a subset of mechanisms will be insufficient to either “rule out” any policies, or to identify policies that are worth doing or at least testing and considering further. This contrast is highlighted across the rows of [Table 1](#). The other relevant dimension that varies across the “cells” of [Table 1](#) is the cost or feasibility of carrying out a policy evaluation, which we would always wish to do (regardless of what we had learned from a mechanism experiment) were it cost-less to do so but sometimes is very costly or even impossible.

This framework suggests that under a very particular set of conditions, mechanism experiments by themselves may be sufficient to inform policy decisions. Probably more common are situations in which mechanism experiments and more traditional policy evaluations (which could be either randomized or “natural” experiments) are complements. Under some circumstances, mechanism experiments may not be that helpful and it may be most productive to just go right to running a “black-box” policy evaluation. In this section, we discuss the conditions under which mechanism experiments and policy evaluations will be substitutes and those where they will be complements, and illustrate our key points and the potential scientific and policy impact using different studies that have been carried out.

## 4.1 Mechanism experiment is sufficient

Mechanism experiments alone may be sufficient to guide policy decisions when economists have some prior beliefs about the candidate mechanisms through which a policy might affect outcomes (and so can design relevant mechanism experiments), while testing the real-world policy lever of ultimate interest is impossible—or at least would entail extraordinarily high cost. Under those conditions, a mechanism experiment could be enough to inform a policy decision if there is just a single mechanism or at least a relatively short list of mechanisms through which the policy may affect outcomes.

If the list of candidate mechanisms through which a policy affects outcomes is “too long” then the only way mechanism experiments could by themselves guide policy would be if we were willing to impose the assumption that the different candidate mechanisms do not have interactive effects. Without this “noninteracting” assumption, a test of one or a subset of candidate mechanisms would not tell us anything of much value for policy since there would always be the possibility that implementing the policy that activated the full menu of mechanisms could have much bigger (or much smaller) effects because of the possibility of interactions among the mechanisms. This condition is likely to be quite rare in practice and so in what follows we focus instead on discussing scenarios under which there is just one mechanism, or there are multiple mechanisms (but not too many of them) that could have interactive effects.

### 4.1.1 A single mechanism

One scenario under which a mechanism experiment might be enough to guide policy is when there is just a single mechanism ( $M$ ) that links the candidate policy ( $P$ ) to the outcome(s) of policy concern ( $Y$ ). A mechanism experiment is most likely to be sufficient for this purpose if we already understand something about the causal link that carries the policy to the outcome; that is, if we already know either the effect of the policy on the mechanism ( $P \rightarrow M$ ), and so just need to learn more about the effects of the mechanism on the outcome ( $M \rightarrow Y$ ), or vice versa.

Consider an example from the area of education policy. A key goal of many public policies is to promote college attendance, particularly among low-income people, as a way to achieve redistributional goals and account for positive externalities from schooling attainment. An important open question is the degree to which low levels of college attendance by low-income people is due to the price of college versus the effect of poverty on academic preparation, that is, on how much people learn over their elementary and secondary school careers and so how ready they are to do college-level work. Policies to reduce the price of college among low-income potential college-goers include federal financial aid, especially Pell grants. The existing evidence on the link between financial aid and college attendance has been mixed. Some state-level programs appear to have had large effects, while others have not. In nonexperimental studies, the effects of national changes in the Pell grant program itself have been difficult to disentangle from national changes in other factors affecting college attendance.

This is an example where a mechanism experiment might be enough to guide policy. The candidate mechanism of interest here is price ( $M$ ), and the key policy question is the degree to which the price of college affects attendance and completion ( $M \rightarrow Y$ ). Providing additional financial aid lowers the price ( $P \rightarrow M$ ); there are of course questions about the exact magnitude of that relationship and who the “compliers” would be with any given policy change, but at least we can sign that effect. The key puzzle then is to understand the  $M \rightarrow Y$  link. A mechanism experiment can then test this link, while leaving other aspects of the policy environment, such as the “offer” implied by underlying Pell eligibility criteria and the information provided about college that one receives through the Pell application process, as fixed.

The study by [Bettinger et al. \(2012\)](#) builds on the insight that if the key candidate mechanism through which efforts to change educational attainment is the price of college, then potentially *any* intervention that changes this mechanism can provide useful information about the effects of college price on college attendance or persistence (that is, on the  $M \rightarrow Y$  link).<sup>11</sup> Their study generates useful information about the potential

<sup>11</sup> We say “potentially” here because there is a key assumption here about whether the nature of the  $M \rightarrow Y$  link depends on the specific  $P$  that is used to modify the value of  $M$ ; we discuss this issue in greater detail below.

effects of changes to the large-scale Pell grant program by testing an intervention that looks like a change to Pell grant generosity—specifically, the authors worked with H&R Block to increase *take-up* of the existing Pell grants and other federal financial aid programs through the personal assistance of a tax preparer. Note that any other intervention that changed federal financial aid take-up could also have been used. But this particular experiment employed a narrowly targeted form of outreach to customers of tax preparers about whom much financial information was known, and thus probably had much lower costs per additional person aided than broader types of advertising and outreach would.<sup>12</sup>

There are few mechanisms through which the H&R Block intervention might plausibly affect college attendance *besides* receipt of financial aid itself. The most likely alternative mechanism is the possibility of increased general awareness of college and its costs. To examine the empirical importance of this second candidate mechanism, the researchers added a second arm to the experiment which tested the effects of additional general information about college.

The magnitude of the change caused in financial aid received was substantial. For instance, among dependent children whose families received the personal assistance in the experiment, aid increased from \$2360 to \$3126, on average. This mechanism experiment found that college attendance increased from 28% to 36% among high school seniors whose parents received the personal assistance, and the outcomes of people receiving only additional information were unaffected. We interpret the results as consistent with the idea that the price of college is an important factor determining college attendance, for at least a subset of low-income people; that is, at relatively low cost, we have documented the magnitude of the  $M \rightarrow Y$  link. Ideally, we would also do a policy evaluation to better understand take-up rates and the overall magnitude for the change in college price that would result from changing Pell grant generosity, but because the  $P \rightarrow M$  link is better understood than the  $M \rightarrow Y$  link, the mechanism experiment can be combined with that prior knowledge to generate some additional useful information for policy.

Now consider a different example from the area of urban policy that helps highlight some of the additional assumptions that might be required to rely just on a mechanism experiment to guide policy. A key concern for many cities in the United States is the potential adverse effects on health in high-poverty neighborhoods from the limited availability of grocery stores—so-called “food deserts.” The actual policy intervention that is often considered as a response to this potential problem is to subsidize grocery stores to locate into disadvantaged communities. Carrying out a policy evaluation of

<sup>12</sup> Of course, a different policy lever that could be used here is simplification of the process for applying for financial aid, which could potentially also be done at low cost. But a test of this policy change, as with a direct test of changing the Pell grant generosity itself, could only be accomplished through changes in laws.

location incentives for grocery stores would be very costly because the unit of randomization would be the community, the cost per community is high, and the number of communities needed to provide adequate statistical power to detect impacts is large.

The possibility of using a lower-cost mechanism experiment to understand the value of this intervention stems from the plausible assumption that changes in eating healthy foods (fruits, vegetables, whole grains) is the key mechanism ( $M$ ) through which introducing grocery stores into high-poverty urban areas would improve health, and the recognition that previous research tells us something about the effects of eating healthy foods on health—that is, we already know the  $M \rightarrow Y$  link. Consider the following mechanism experiment that could be carried out instead: Enroll a sample of low-income families, and randomly assign some of them (but not others) to receive free weekly delivery of fresh fruits and vegetables to their homes. By using individuals (rather than communities) as the unit of randomization, this mechanism experiment would be much less expensive than a policy evaluation of the actual policy of interest (subsidized grocery store location). The reduction in costs associated with randomizing people rather than neighborhoods also lets us test a “treatment dose” that is much more intensive than what could be obtained with any realistic policy intervention.

Imagine we found that several hundreds of dollars’ worth of free fruits and vegetables delivered to someone’s door each month had *no effect* on obesity. This would tell us that even though healthy eating ( $M$ ) has important impacts on health ( $Y$ ), changing eating habits ( $M$ ) through even fairly intensive interventions ( $P$ ) is challenging in practice. The set of policies about which we could draw conclusions from this mechanism experiment would depend on how much we believe we know about the nature of the  $P \rightarrow M$  link. Suppose we also believed eating habits adapt rapidly to changes in food availability, that social interactions are not very important in shaping eating habits, and that reducing the price of accessing healthy food never *reduces* the chances of eating them (that is, there is a monotonic relationship between the treatment dose and the treatment response); in that case, null results from our mechanism experiment would lead us to predict that *any* sort of policy that tried to address the “food desert” problem would (on its own) be unlikely to diminish problems related to obesity.

If we had more uncertainty about the role of social interactions or time in affecting eating habits, then different mechanism-experiment designs would be required. If we believed that social interactions might be important determinants of people’s eating habits, then we would need a more costly experiment with three randomized arms, not just two—a control group, a treatment arm that received free food delivery for themselves, and a treatment arm that received food delivery for themselves and for a limited

number of other households that the family designated (“buddy deliveries”).<sup>13</sup> If we thought that eating habits were determined at a still larger macro-level, we would have to randomly assign entire communities to receive free home food delivery. A community-level test of home fruit and vegetable delivery could still wind up being less expensive than a policy evaluation of incentive locations for grocery stores, because of the large guarantees that would be required to entice a grocery store to incur the start-up costs of establishing a new location in a neighborhood. But if we thought that eating habits changed very slowly over time, and at the community level, then we would have to commit to providing home food delivery for entire communities for extended periods of time—at which point there might be little cost advantage compared to a policy evaluation of grocery-store subsidies.

#### **4.1.2 Multiple (but not too many) candidate mechanisms**

In some situations, it may be possible to learn about the effects of a policy without ever doing a policy evaluation, so long as the list of candidate mechanisms is not “too long.” In this case, mechanism experiments can still turn out to be lower-cost ways of generating the necessary policy-relevant information compared to carrying out a full-blown policy evaluation.

Consider a policy (P) that may affect some outcome (Y) through three different candidate mechanisms, given by  $M_1$ ,  $M_2$ , and  $M_3$ . If these mechanisms could potentially have interactive effects—that is, the different mechanisms could either amplify or undercut each other’s effects—then in a world without resource or feasibility constraints, clearly the best way to test the net effect of the policy would be to carry out a policy evaluation. But sometimes policy evaluations are not feasible, or even if they are, they are enormously costly. In some circumstances, it may be possible to learn about the effect of the policy at lower cost through a mechanism experiment that reduces the cost of learning about at least some of the mechanisms and their interactions with the other mechanisms through interventions that do not look like the policy of interest.

For example, one way to do this is by avoiding the cost of implementing one of the mechanisms (say,  $M_1$ ) by exploiting naturally occurring population variation in that factor to understand interactivity with the other candidate mechanisms ( $M_2$  and  $M_3$ ). As an illustration of this idea, consider one of the “kitchen-sink” policy evaluations of the sort that the federal government sometimes supports, like Jobs Plus. This experiment

<sup>13</sup> Duflo and Saez (2003) discuss a cleverly designed experiment that used individuals as the unit of analysis but was designed to identify spillover effects. In their experiment, some people in some departments within a company received incentives to visit a benefit fair to learn more about savings plans. They assessed both direct effects of the information, and effects of information spillovers (from comparisons of the outcomes of the non-incentivized individuals in incentivized departments to individuals in non-incentivized departments). The information diffused through the experiment had a noticeable impact on plan participation.

tested the combined effects of providing public housing residents with financial incentives for work (relief from the “HUD tax” on earnings that comes from setting rent contributions as a fixed share of income—call this  $M_1$ ), employment and training services ( $M_2$ ), and efforts to improve “community support for work” ( $M_3$ ). Previous studies have already examined the effects of the first two program ingredients when administered independently, while the potential value of community support for work is suggested by the work of sociologist [William Julius Wilson \(1987, 1997\)](#) among others. The key program theory of Jobs Plus is that these three mechanisms interact and so have more-than-additive effects on labor market outcomes ([Bloom et al., 2005](#)), so carrying out three separate experimental tests of each independent mechanism would obviously not be informative about what would result from the full package. So the bundle was tested with a policy evaluation carried out across six cities, in which entire housing projects were randomly assigned to either a control group or a program group in which residents received the bundle of Jobs Plus services.

What would a lower-cost mechanism experiment look like in this case? Imagine enrolling people who are already living in neighborhoods with high employment rates—so that there is already “community support for work” ( $M_3$ ) in place “for free” to the researchers. This already makes the intervention being tested look quite different from the actual policy of interest, since the policy is motivated by concern about helping a population that is exactly the opposite of the one we would be targeting—that is, the policy wants to help people in areas with *low* employment rates. Such a design would allow us to capture the interaction of community support for work with other aspects of the policy, although not its main effect. Suppose within these we identify people receiving means-tested housing assistance in those areas, then we randomly assign some of them to receive no reduction in benefits as their income rose ( $M_1$ ) and employment and training services ( $M_2$ ).

Our proposed mechanism experiment conserves resources by reducing the dimensionality of the experimental intervention. If we did find some evidence of an effect using this design, we could carry out a follow-up mechanism experiment that included people living in both high- and low-employment neighborhoods—this would let us see how varying the value of  $M_3$  changes the effects of varying the value of the other two mechanisms. This variation in the mechanism is obviously nonexperimental; whether this series of mechanism experiments would dominate just carrying out a full-blown policy evaluation of Jobs Plus would depend partly on how we viewed the trade-off between some additional uncertainty versus additional research costs.

## 4.2 Mechanism experiment plus policy evaluation

In this section, we discuss different scenarios under which it makes sense to carry out both mechanism experiments and policy evaluations, and provide some examples from

previous research. We begin by discussing scenarios in which the mechanism experimentation would come first followed by a policy evaluation, and then scenarios under which the optimal sequence would likely be reversed. Note that even when a mechanism experiment has to be followed by a policy evaluation, the mechanism experiment may still add value by helping us figure out which evaluations are worth running. This includes carrying out mechanism experiments in different settings to determine *where* it is worth trying a policy evaluation.

#### **4.2.1 Mechanism experiments then policy evaluation**

Mechanism experiments can help concentrate resources on testing part of a causal chain that links a policy to an outcome. One reason it would make sense to follow a mechanism experiment that had encouraging results with a full-blown policy evaluation would be to learn more about the other parts of the causal chain. An example would be a mechanism experiment that documents that a given mechanism affects some outcome of policy concern ( $M \rightarrow Y$ ), but now for policy purposes we need to also understand the other part of the chain ( $P \rightarrow M$ ). The mechanism experiment can add value here by identifying those applications where the mechanism is unrelated to the outcome ( $M \rightarrow Y = 0$ ) and so avoiding the need to carry out a costly follow-up policy evaluation.

For example, we have argued above that the low-poverty voucher treatment within the MTO residential-mobility demonstration can be thought of as a mechanism experiment—it tests an intervention causing moves to lower poverty areas that is unlikely to ever be implemented as a real policy. This treatment arm makes clear that living in a low-poverty neighborhood of the sort that families with regular housing vouchers move into on their own can have beneficial effects for physical and mental health, delinquency and perhaps even for children’s long-term earnings prospects during adulthood. This finding motivates follow-up policy evaluations that test more realistic changes to the voucher policy that might also help steer families into lower poverty areas without an (unrealistic) mandate. Such policies include more intensive mobility counseling or supports compared to what was provided in MTO, or changes in the voucher program design that increases subsidy amounts in lower poverty areas (Collinson and Ganong, 2014).

A different scenario under which it may be worth following a mechanism experiment with a policy evaluation is when implementation of a policy is a significant factor in the causal chain. Medical researchers distinguish between “efficacy trials,” which are small-scale research trials of model programs carried out with high fidelity, and “effectiveness trials” that test the effects of some intervention carried out under field conditions at scale. Efficacy trials can be thought of as a type of mechanism experiment. Compared to efficacy trials, effectiveness trials often have more program attrition, weaker training of service providers, weaker implementation monitoring, and smaller impacts

([Lipsey et al., 2007](#)). Thus, an efficacy trial may test the effect of a high-fidelity treatment on a health outcome ( $M \rightarrow Y$ ) and an effectiveness trial may show the effect of a policy on provider implementation ( $P \rightarrow M$ ) as well as the overall effect of a lower fidelity treatment on health ( $P \rightarrow Y$ ).

Sometimes, mechanism experiments can also help highlight lower cost interventions to test with subsequent policy evaluations. Imagine a situation in which we have a policy  $P_0$  attempting to achieve outcome  $Y$ , and that  $P_0$  may work through numerous mechanisms  $M_1 \dots M_n$ . If a mechanism experiment found a strong effect of  $M_1$  on the outcome, it may be possible to design a simpler policy  $P_1$  that works only through  $M_1$  and is less expensive.

Consider as an example policies that are summer interventions to address the challenges that children from low-income families face maintaining academic gains over the summer. There is a great deal of concern about summer learning loss among poor children relative to more affluent ones. It has long been hypothesized that the loss is due to more limited involvement with academically or cognitively stimulating activities over the summer ([Alexander et al., 2007](#)). Potential policy interventions that have been implemented or proposed are to subsidize summer programming for youth ([Fifer and Krueger, 2006](#)). To the extent that these interventions look like summer school, they are expensive like summer school.

In this context, we could consider the study by [Guryan et al. \(2014\)](#) as a mechanism experiment that tests one candidate mechanism through which summer school might improve academic outcomes—by increasing the amount of reading students do over the summer. Their study tests this mechanism by sending books directly to the homes of low-income children. The results of that experiment find substantial impacts on reading scores for some students later into the academic year. The implication is that a “summer books” intervention could potentially turn out to be even more cost-effective than summer school itself, and so might warrant a large-scale policy evaluation to calibrate magnitudes.

Since mechanism experiments test an isolated segment of a causal chain, a natural question in this case is to wonder why we do not just test the other parts of the causal chain using separate mechanism experiments. In many cases that might be possible. But one subtle reason this might not work, and so why a follow-up policy evaluation would be required, would be if the link between the mechanism and the outcome ( $M \rightarrow Y$ ) depends on the specific policy lever ( $P$ ) that is used. That is, the ( $M \rightarrow Y$ ) link might not be what John DiNardo terms “nonimplementation specific” or what [Heckman \(2010\)](#) calls “policy invariant.” In some situations, it might be possible to determine that the ( $M \rightarrow Y$ ) link is unlikely to be policy invariant by estimating that relationship in several different mechanisms that manipulate the value of  $M$  through some intervention ( $P$ ) other than the true policy of interest. But in other applications, there may be no substitute for

understanding the  $(M \rightarrow Y)$  link when  $M$  is manipulated by the actual policy being considered—that is, to do a policy evaluation.<sup>14</sup>

Some simple notation helps illustrate the problem. Let  $P$  be the policy,  $M$  be the mediator,  $Y$  be the outcome (with  $P \rightarrow M \rightarrow Y$  as in Fig. 1), with  $M = U + V$ ,  $\text{cov}(U, V) = 0$ ,  $\text{cov}(U, Y) = 0$ , and  $\text{cov}(V, Y) > 0$ . That is, only the  $V$  part of  $M$  is causally related to  $Y$ . In population data, we see  $\text{cov}(M, Y) > 0$ . In this example,  $M$  is an implementation specific mediator because policies that change the  $V$  part of  $M$  will change  $Y$ , but policies that change only the  $U$  part of  $M$  will not influence  $Y$ .<sup>15</sup>

#### **4.2.2 Policy evaluation followed by mechanism experiment**

The same logic and potential gains come from a mechanism experiment that documents the effects of a policy on some mechanism ( $P \rightarrow M$ ), followed by a policy evaluation that helps fill in the effects of the mechanism on the outcome of policy concern ( $M \rightarrow Y$ ).

Consider an example from social policy efforts to improve the long-term life outcomes of disadvantaged youth. Recognizing that youth have multiple needs, many interventions in this area bundle together different intervention elements into a single social program. One example of this is the *Becoming a Man* (BAM) intervention, which was designed by Chicago-area nonprofit *Youth Guidance*. BAM is an in-school intervention delivered to youth in groups that uses principals of cognitive behavioral therapy (CBT) to try to get youth to recognize situations in which their automatic responses (what psychologists call “system 1”; see for example Kahneman, 2011) may get them into trouble, and slow down and be more reflective (“system 2”) before they act. There are some other candidate mechanisms through which BAM might change youth behavior (such as changes in self-control or “grit”) that can largely be ruled out through surveys that the Chicago Public Schools administered to youth in both the treatment and control groups.<sup>16</sup> Yet, one candidate mechanism that many practitioners believe to be quite important for all youth programs is simply putting youth into regular contact with prosocial adults—that is, a basic “mentoring” effect.

A policy evaluation of BAM as implemented in the 2009–10 academic year found the intervention reduces violent-crime arrests by 44% of the mean rate estimated for

<sup>14</sup> One reason we might not see policy invariance is if there is treatment effect heterogeneity in how people’s outcomes respond to some mechanism and people also vary in how the value of that mechanism responds to a change in a policy. In this case, who specifically the “compliers” are whose value of  $M$  is induced to change by a given  $P$  will play an important role in determining what the ultimate effect of the policy is on the outcomes of interest ( $P \rightarrow Y$ ). As a real-world example, consider the case of mental health parity for health insurance. Efficacy trials in medicine are able to establish that certain types of mental health treatment improve mental health outcomes. But the effect of the policy on population mental health will depend critically on who the compliers are—the people who whose mental health treatment status changed by the law.

<sup>15</sup> Our thanks to Steve Pischke for this suggestion.

<sup>16</sup> Administrative data rule out the idea that incapacitation is an important effect (since reductions in arrests are not limited to those days when after-school programming is in session).

people who would have participated in the intervention if it had been offered to them (the control complier mean), while a follow-up study in 2013–15 that found similarly large reductions (see Heller et al., 2013; Heller et al., 2015).

Whether the effect of BAM is due to the CBT curriculum itself or instead to a generic “mentoring effect” is of some policy relevance. If the effect were due merely to exposure to a prosocial adult, then any number of nonprofits in Chicago (or anywhere around the country) could be enlisted to provide services to youth, since there would be nothing specific to the BAM curriculum or how it is delivered that would matter. On the other hand, if the content of the CBT curriculum is important, then efforts to deliver that content with fidelity becomes critical. One could imagine following up the BAM policy evaluations with two different types of mechanism experiments. One might have some BAM counselors run versions of the program that essentially threw the curriculum out the window (the weekly group meetings of the youth would be unstructured and just focus on building rapport between the counselors and the youth), or even have youth engage in peer-to-peer mentoring. The other mechanism experiment might (say) enroll youth in the equivalent of an online CBT massive online open course (MOOC) so that there would be no new connection created to any prosocial adult. While these mechanism experiments would have benefits for policy design, it is at the same time not hard to imagine how policymakers (and nonprofit providers) might have had the initial reaction of objecting to the idea of testing what would feel like ‘watered-down’ versions of BAM that eliminated either the curriculum or the connection to the counselor.

### 4.3 Just do policy evaluation

A scenario under which the most productive strategy may be to just do a policy evaluation is one in which the policy of interest has a long list of candidate mechanisms that could have interactive effects or work at cross-purposes. Under that type of circumstance, the number of mechanism experiments that would be needed to test different combinations of candidate mechanisms would be large, and because of the possibility of interactive effects it may ultimately require a treatment arm that included all candidate mechanisms. At that point, there is no cost advantage from preceding a policy evaluation with a mechanism experiment—researchers should just go straight to doing a policy evaluation.

Consider for example the effects of changing police staffing levels on crime rates. This is an important policy question because the United States spends over \$100 billion per year on police,<sup>17</sup> and hiring more police is an extremely scalable intervention—the one thing that almost every police department in the country can do consistently at large

<sup>17</sup> The figure in 2006 was \$99 billion <http://www.albany.edu/sourcebook/pdf/t122006.pdf>.

scale. Moreover, there remains great debate within the social science community about whether simply putting more police on the street will reduce crime, with most economists of the view that it will while conventional wisdom within criminology remains largely skeptical.

Above, we illustrated the potential value of using mechanism experiments to reduce the costs of understanding treatment effect heterogeneity (by narrowing the set of contexts in which we would need to carry out a policy evaluation) by focusing on a single mechanism through which stepped-up police staffing might affect crime is by changing victim reporting to the police. But in reality, there are many other potential channels as well; for example, police may incapacitate offenders even without victim reporting if police happen upon a crime that occurs in the act. Police presence itself could also directly deter crime, even aside from victims calling the police to report crimes. On the other hand, putting more police on the street could potentially have adverse effects on crime if the result is to exacerbate police-community tensions, or if policing is carried out in a way that reduces perceived legitimacy of the law and the criminal justice system, or if the incapacitation effects of policing are actually negative—that is, if putting more people in jail or prison weakens communities and suppresses informal sources of social control. Understanding the effects of just a subset of these mechanisms would inevitably leave open the key question for policy, which about the net effect of the full bundle of mechanisms that come from putting more police on the street.

The best strategy in this case would be to simply carry out a policy evaluation of what happens from putting more police in some areas but not others. This has been the topic of a large body of work within criminology, in which police departments working with researchers randomly assign extra patrol activity to some high crime “hot spots” but not others; see for example Braga et al. (2012). The one challenge in that literature comes from the possibility of general equilibrium or spillover effects—that is, the possibility that saturating some areas with police could lead criminals to migrate to other areas, or what criminologists call “displacement.” In principle, one solution to that problem would be to just carry out random assignment at increasingly large geographic levels. In practice, economists have overcome this problem by relying on natural experiment variation instead (e.g., Evans and Owens, 2007).

A different scenario under which it makes sense to just carry out a policy evaluation directly, without any preceding mechanism experiments, is when the costs of carrying out policy evaluations are very low. This often arises in practice in situations where there is some government service for which there is excess demand, and policymakers use random lotteries as a rationing device. Examples include charter schools or magnet schools, which in many cities and states must use admissions lotteries as a matter of law (see for example Cullen et al., 2006), low-income housing programs, which at present are funded at a level that enables fewer than one-in-four income-eligible households to participate and so leads many cities to use lotteries (see for example Jacob and

Ludwig, 2012; Jacob et al., 2015), and the expansion of Medicaid in Oregon in 2008 (see Taubman et al., 2014; Baicker et al., 2014, 2013; Finkelstein et al., 2012). In our view, randomized lotteries conducted by governments to provide fair access to programs can be turned into field experiments with the appropriate collection of data about all participants in the lottery, regardless of the lottery's outcome.

## 5. CONCLUSION

In the area of social policy, a great deal of field experimentation is ultimately in the service of informing policy design. If we change the incentives of students and teachers, can we learn how to operate schools to get better educational outcomes? If we vary the structure of health insurance marketplaces, can we learn about how beneficiaries make choices in a way that will allow us to promote broader and cheaper coverage? Questions such as these are at the heart of the movement toward greater use of experimental evidence for social policy design.

The value of a well-executed field experiment is the claim to internal validity—that is, the claim that we have learned something about the effects of the policy of interest in the context in which the policy was tested in the experiment. However, policymakers are often responsible for making decisions about a wide range of contexts beyond those studied in any given policy evaluation. Abstracting from budgetary or feasibility constraints, experimental methods in the form of policy evaluations carried out in different policy-relevant contexts can answer the key questions of policy interest by testing a proposed policy directly. But, in reality, researchers and policymakers alike do in fact face those constraints.

What we have argued in this chapter is that, under some circumstances, the most efficient way to learn about the effectiveness of a policy is not always a direct test of the policy; in fact, what can be most useful are field experiments that bear little surface resemblance at all to the policy of interest. When we have information or beliefs about the *mechanisms* by which policies operate, we can sometimes generate more policy-relevant information per dollar spent by carrying out a mechanism experiment instead of a policy evaluation, and mechanism experiments can sometimes also help improve our forecasts for the contexts under which a policy would be expected to have effects.

Ultimately, then, for researchers and policymakers the issue becomes one of problem selection—what, precisely, should we seek to use field experiments to test? In our view, the portfolio of field experiments in the area of social policy should not consist entirely of mechanism experiments. Policy evaluations will always play a critical role, but there is currently so little attention to mechanism experiments designed to inform policy questions that there may be considerable value in expanding the use of them in practice.

## REFERENCES

- Alatas, V., Banerjee, A., Hanna, R., Olken, B.A., Purnamasari, R., Wai-Poi, M., 2013. Ordeal Mechanisms in Targeting: Theory and Evidence from a Field Experiment in Indonesia. Working Paper 19127. National Bureau of Economic Research. <http://www.nber.org/papers/w19127>.
- Alexander, K.L., Entwistle, D.R., Olson, L.S., 2007. Lasting consequences of the summer learning gap. *Am. Sociol. Rev.* 72 (2), 167–180.
- Angrist, J.D., Pischke, J.-S., 2009. *Mostly Harmless Econometrics*. Princeton University Press, Princeton, NJ.
- Angrist, J.D., Pischke, J.-S., 2010. The credibility revolution in empirical economics: how better research design is taking the con out of econometrics. *J. Econ. Perspect.* 24 (2), 3–30.
- Ashraf, N., Berry, J., Shapiro, J.M., 2010. Can higher prices stimulate product use? Evidence from a field experiment in Zambia. *Am. Econ. Rev.* 100 (5), 2383–2413. <http://dx.doi.org/10.1257/aer.100.5.2383>.
- Baicker, K., Taubman, S., Allen, H., Bernstein, M., Gruber, J., Newhouse, J.P., Schneider, E., Wright, B., Zaslavsky, A., Finkelstein, A., The Oregon Health Study Group, 2013. The Oregon experiment – effects of medicaid on clinical outcomes. *N. Engl. J. Med.* 368 (18), 1713–1722.
- Baicker, K., Finkelstein, A., Song, J., Taubman, S., 2014. The impact of medicaid on labor market activity and program participation: evidence from the Oregon health insurance experiment. *Am. Econ. Rev. Pap. Proc.* 104 (5), 322–328.
- Banerjee, A.V., Duflo, E., 2009. The experimental approach to development economics. *Annu. Rev. Econ.* 1, 151–178.
- Bellisle, D., Marzahl, D., September 2015. Restructuring the EITC: A Credit for the Modern Worker. Center for Economic Progress, Chicago, IL.
- Bettinger, E.P., Terry Long, B., Oreopoulos, P., Sanbonmatsu, L., 2012. The role of application assistance and information in college decisions: results from the H&R block FAFSA experiment. *Q. J. Econ.* 127 (3), 1205–1242.
- Bhargava, S., Manoli, D., 2013. Why Are Benefits Left on the Table? Assessing the Role of Information, Complexity, and Stigma on Take-up with an IRS Field Experiment. Unpublished Working Paper.
- Bloom, H.S., Orr, L.L., Bell, S.H., Cave, G., Doolittle, F., Lin, W., Bos, J.M., 1997. The benefits and costs of JTPA title II-a programs: key findings from the national job training partnership act study. *J. Hum. Resour.* 32 (3), 549–576.
- Bloom, H.S., Riccio, J.A., Verma, N., 2005. *Promoting Work in Public Housing: The Effectiveness of Jobs-Plus*. MDRC, New York.
- Bloom, H., Schwartz, S., Lui-Gurr, S., Lee, S.-W., 1999. Testing a re-employment incentive for displaced workers: the earnings supplement project. *Soc. Res. Demonstr. Corp.* <http://www.srdc.org/media/195754/testing.pdf>.
- Braga, A., Papachristos, A., Hureau, D., 2012. Hot spot policing effects on crime. *Campbell Syst. Rev.* 2012, 8.
- Buka, S.L., Stichick, T.L., Birdthistle, I., Earls, F.J., 2001. Youth exposure to violence: prevalence, risks, and consequences. *Am. J. Orthopsychiatr.* 71 (3), 298–310.
- Chetty, R., Hendren, N., Katz, L.F., 2016. The effects of exposure to better neighborhoods on children: new evidence from the moving to opportunity experiment. *Am. Econ. Rev.* 106 (4), 855–902.
- Chetty, R., Friedman, J.N., Saez, E., 2013. Using differences in knowledge across neighborhoods to uncover the impacts of the EITC on earnings. *Am. Econ. Rev.* 103 (7), 2683–2721. <http://dx.doi.org/10.1257/aer.103.7.2683>.
- Cohen, J., Dupas, P., 2010. Free distribution or cost-sharing? Evidence from a randomized malaria prevention experiment. *Q. J. Econ.* 125 (1), 1–45. <http://dx.doi.org/10.1162/qjec.2010.125.1.1>.
- Cole, S.R., Stuart, E.A., 2010. Generalizing evidence from randomized clinical trials to target populations: the ACTG 320 trial. *Am. J. Epidemiol.* 172 (1), 107–115.
- Collinson, R.A., Ganong, P., 2014. The Incidence of Housing Voucher Generosity. Working Paper. <http://ssrn.com/abstract=2255799>.
- Cook, T.D., Campbell, D.T., 1979. Quasi-Experimentation: Design and Analysis Issues for Field Settings. Wadsworth.
- Cullen, J.B., Jacob, B.A., Levitt, S., 2006. The effect of school choice on participants from randomized lotteries. *Econometrica* 74 (5), 1191–1230.

- Davidson, C., 1995. Wage Subsidies for Dislocated Workers, vol. 95, 31. WE Upjohn Institute for Employment Research.
- Deaton, A., 2010. Instruments, randomization, and learning about development. *J. Econ. Lit.* 48 (2), 424–455.
- DiNardo, J., Lee, D.S., 2011. Program evaluation and research designs. In: Ashenfelter, O., Card, D. (Eds.), *Handbook of Labor Economics*, vol. 4, Part A. Elsevier, Amsterdam, pp. 463–536.
- Duflo, E., Saez, E., 2003. The role of information and social interactions in retirement plan decisions: evidence from a randomized experiment. *Q. J. Econ.* 118 (3), 815–842.
- Evans, W.N., Owens, E., 2007. Cops and crime. *J. Public Econ.* 91 (1–2), 181–201.
- Feynman, R., 1964. A video in the Messenger Lecture Series. Quotation starts at 38:48. The Great Conservation Principles. Available at: <http://research.microsoft.com/apps/tools/tuva/index.html#data=4|84edf183-7993-4b5b-9050-7ea34f236045||>.
- Fifer, M.E., Krueger, A.B., 2006. Summer Opportunity Scholarships: A Proposal to Narrow the Skills Gap. 2006–03. Hamilton Project Discussion Paper. [http://www.brook.edu/views/papers/200604hamilton\\_3.pdf](http://www.brook.edu/views/papers/200604hamilton_3.pdf).
- Figlio, D.N., Rouse, C.E., 2006. Do accountability and voucher threats improve low-performing schools? *J. Public Econ.* 90 (1), 239–255.
- Finkelstein, A., Taubman, S., Wright, B., Bernstein, M., Gruber, J., Newhouse, J.P., Allen, H., Baicker, K., The Oregon Health Study Group, 2012. The Oregon health insurance experiment. evidence from the first year. *Q. J. Econ.* 127 (3), 1057–1106.
- Fortson, J.G., Sanbonmatsu, L., 2010. Child health and neighborhood conditions: results from a randomized housing voucher experiment. *J. Hum. Resour.* 45 (4), 840–864.
- Fryer Jr, R.G., Levitt, S.D., 2004. Understanding the black-white test score gap in the first two years of school. *Rev. Econ. Stat.* 86 (2), 447–464.
- Greenberg, D., Shroder, M., 2004. *The Digest of Social Experiments*, third ed. Urban Institute Press, Washington, DC.
- Gueron, J.M., Rolston, H., 2013. *Fighting for Reliable Evidence*. Russell Sage Foundation.
- Guryan, J., Kim, J.S., Quinn, D.M., 2014. Does Reading During the Summer Build Reading Skills? Evidence from a Randomized Experiment in 463 Classrooms. Working Paper 20689. National Bureau of Economic Research. <http://www.nber.org/papers/w20689>.
- Hastings, J.S., Weinstein, J.M., 2008. Information, school choice, and academic achievement: evidence from two experiments. *Q. J. Econ.* 123 (4), 1373–1414.
- Harris, J.E., 1985. Macro-experiments versus micro-experiments for health policy. In: Hausman, J., Wise, D. (Eds.), *Social Experimentation*. University of Chicago Press, Chicago, pp. 145–185.
- Harrison, G.W., List, J.A., 2004. Field experiments. *J. Econ. Lit.* 42 (4), 1009–1055.
- Hausman, J.A., Wise, D.A., 1985. *Social Experimentation*. University of Chicago Press, Chicago.
- Heckman, J.J., 2010. Building bridges between structural and program evaluation approaches to evaluating policy. *J. Econ. Lit.* 48 (2), 356–398.
- Heckman, J.J., Humphries, J.E., Mader, N., 2011. The GED. In: Hanushek, E.A., Machin, S., Woßmann, L. (Eds.), *Handbook of the Economics of Education*, vol. 3. North Holland, Elsevier, Amsterdam, pp. 423–484 (Chapter 9).
- Heller, S.B., Pollack, H.A., Ander, R., Ludwig, J., 2013. Preventing Youth Violence and Dropout: A Randomized Field Experiment. NBER Working Paper 19014.
- Heller, S.B., Shah, A.K., Guryan, J., Ludwig, J., Mullainathan, S., Pollack, H.A., 2015. Thinking, Fast and Slow? Some Field Experiments to Reduce Crime and Dropout in Chicago. Working paper.
- Hotz, V.J., Imbens, G.W., Mortimer, J.H., 2005. Predicting the efficacy of future training programs using past experiences at other locations. *J. Econ.* 125 (1–2), 241–270.
- Hoxby, C.M., 2003. School choice and school productivity: could school choice be a tide that lifts all boats? In: Hoxby, C.M. (Ed.), *The Economics of School Choice*. University of Chicago Press, pp. 287–342.
- Imbens, G.S., 2010. Better late than nothing: some comments on Deaton (2009) and Heckman and Urzua (2009). *J. Econ. Lit.* 48 (2), 399–423.

- Jacob, B.A., Ludwig, J., 2012. The effects of housing assistance on labor supply: evidence from a voucher lottery. *Am. Econ. Rev.* 102 (1), 272–304.
- Jacob, B.A., Kapustin, M., Ludwig, J., 2015. The impact of housing assistance on child outcomes: evidence from a randomized housing lottery. *Q. J. Econ.* 130 (1).
- Jones, D., 2010. Information, preferences, and public benefit participation: experimental evidence from the advance EITC and 401(k) savings. *Am. Econ. J. Appl. Econ.* 2 (2), 147–163.
- Kabat-Zinn, J., Massion, A.O., Kristeller, J., Peterson, L.G., Fletcher, K.E., Pbert, L., Lenderking, W.R., Santorelli, S.F., 1992. Effectiveness of a meditation-based stress reduction program in the treatment of anxiety disorders. *Am. J. Psychiatr.* 149 (7), 936–943.
- Kahneman, D., 2011. Thinking, Fast and Slow. Macmillan.
- Karlan, D., Zinman, J., 2009. Observing unobservables: identifying information asymmetries with a consumer credit field experiment. *Econometrica* 77 (6), 1993–2008. <http://dx.doi.org/10.3982/ECTA5781>.
- Keizer, K., Lindenberg, S., Steg, L., 2008. The spreading of disorder. *Science* 322 (5908), 1681–1685.
- Kelling, G.L., Wilson, J.Q., March 1982. Broken windows. *Atl. Mon.* <http://www.theatlantic.com/magazine/archive/1982/03/broken-windows/4465/>.
- Kessler, R.C., Duncan, G.J., Gennetian, L.A., Katz, L.F., Kling, J.R., Sampson, N.A., Sanbonmatsu, L., Zaslavsky, A.M., Ludwig, J., 2014. Associations of randomization in a housing-mobility experiment with mental disorders among low-income adolescence. *J. Am. Med. Assoc.* 311 (9), 937–947.
- Kletzer, L.G., Litan, R.E., 2001. A Prescription to Relieve Worker Anxiety. Policy Brief 73. Brookings Institution. <https://www.piie.com/publications/pb/pb.cfm?ResearchID=70>.
- Kling, J.R., 2007. Methodological frontiers of public finance field experiments. *Natl. Tax J.* 60 (1), 109–127.
- Kling, J.R., Liebman, J.B., Katz, L.F., 2007. Experimental analysis of neighborhood effects. *Econometrica* 75 (1), 83–119.
- Kling, J.R., Ludwig, J., Katz, L.F., 2005. Neighborhood effects on crime for female and male youth: evidence from a randomized housing voucher experiment. *Q. J. Econ.* 120 (1), 87–130.
- LaLonde, R.J., 2007. The Case for Wage Insurance. Council Special Report 30. <http://www.cfr.org/world/case-wage-insurance/p13661>.
- Levitt, S.D., 1998. The relationship between crime reporting and police: implications for the use of uniform crime reports. *J. Quant. Criminol.* 14 (1), 61–81.
- Lipsey, M.W., Landenberger, N.A., Wilson, S.J., 2007. Effects of cognitive-behavioral programs for criminal offenders. *Campbell Syst. Rev.*
- Ludwig, J., Duncan, G.J., Gennetian, L.A., Katz, L.F., Kessler, R.C., Kling, J.R., Sanbonmatsu, L., 2013. Long-term neighborhood effects on low-income families: evidence from moving to opportunity. *Am. Econ. Rev. Pap. Proc.* 103 (3), 226–231.
- Ludwig, J., Duncan, G.J., Gennetian, L.A., Katz, L.F., Kessler, R.C., Kling, J.R., Sanbonmatsu, L., 2012. Neighborhood effects on the long-term well-being of low-income adults. *Science* 337 (6101), 1505–1510.
- Ludwig, J., Sanbonmatsu, L., Gennetian, L., Adam, E., Duncan, G.J., Katz, L., Kessler, R., Kling, J., Tessler Lindau, S., Whitaker, R., McDade, T., 2011a. Neighborhoods, obesity, and diabetes—a randomized social experiment. *N. Engl. J. Med.* 365 (16), 1509–1519.
- Ludwig, J., Kling, J.R., Mullainathan, S., 2011b. Mechanism experiments and policy evaluations. *J. Econ. Perspect.* 25 (3), 17–38.
- Ludwig, J., Liebman, J., Kling, J., Duncan, G.J., Katz, L.F., Kessler, R.C., Sanbonmatsu, L., 2008. What can we learn about neighborhood effects from the moving to opportunity experiment? *Am. J. Sociol.* 114 (1), 144–188.
- Manoli, D., Turner, N., 2014. Nudges and learning effects from informational interventions: evidence from notifications for low-income taxpayers. *Natl. Bur. Econ. Res. Working Papers Series*, No. 20718.
- Meyer, B.D., 1995. Natural and quasi-experiments in economics. *J. Bus. Econ. Stat.* 13 (2), 151–161.
- Newhouse, J.P., The Insurance Experiment Group, 1993. Free for All? Lessons from the RAND Health Insurance Experiment. Harvard University Press, Cambridge, MA.

- Poe-Yamagata, E., Jacob Benus, N.B., Hugh Carrington, M.M., Shen, T., 2011. Impact of the reemployment and eligibility assessment (REA) initiative. IMPAQ.
- Phelps, E.S., 1994. Low-wage employment subsidies versus the welfare state. *Am. Econ. Rev.* 84, 54–58.
- Plueger, D., 2009. Earned Income Tax Credit Participation Rate for Tax Year 2005. IRS Research Bulletin.
- Rivera, J.A., Sotres-Alvarez, D., Habicht, J.-P., Shamah, T., Villalpando, S., 2004. Impact of the Mexican program for education, health and nutrition (Progresa) on rates of growth and anemia in infants and young children: a randomized effectiveness study. *JAMA* 291 (21), 2563–2570.
- Rouse, C.E., 1998. Private school vouchers and student achievement: an evaluation of the milwaukee parental choice program. *Q. J. Econ.* 113 (2), 553–602.
- Sanbonmatsu, L., Kling, J.R., Duncan, G.J., Brooks-Gunn, J., 2006. Neighborhoods and academic achievement: results from the moving to opportunity experiment. *J. Hum. Resour.* 41 (4), 649–691.
- Sanbonmatsu, L., Ludwig, J., Katz, L.F., Gennetian, L.A., Duncan, G.J., Kessler, R.C., Adam, E., McDade, T.W., Lindau, S.T., 2011. Moving to Opportunity for Fair Housing Demonstration Program—Final Impacts Evaluation. US Department of Housing & Urban Development, PD&R.
- Schultz, T.P., 2004. School subsidies for the poor: evaluating the Mexican Progresa poverty program. *J. Dev. Econ.* 74 (1), 199–250.
- Sharkey, P., 2010. The acute effect of local homicides on children's cognitive performance. *Proc. Natl. Acad. Sci.* 107 (26), 11733–11738.
- Skoufias, E., Parker, S.W., Behrman, J.R., Pessino, C., 2001. Conditional cash transfers and their impact on child work and schooling: evidence from the PROGRESA program in Mexico. *Economia* 2 (1), 45–96.
- Stuart, E.A., Cole, S.R., Bradshaw, C.P., Leaf, P.J., 2011. The use of propensity scores to assess the generalizability of results from randomized trials. *J. R. Stat. Soc. Ser. A* 174 (2), 369–386.
- Taubman, S., Allen, H., Wright, B., Baicker, K., Finkelstein, A., The Oregon Health Insurance Group, 2014. Medicaid increases emergency department use: evidence from Oregon's health insurance experiment. *Science* 343 (6168), 263–268.
- Todd, P.E., Wolpin, K.I., 2008. Ex ante evaluation of social programs. *Ann. Econ. Stat.* 91/92, 263–292.
- US Department of Education, 2015. FY 2016 Department of Education Justifications of Appropriation Estimates to the Congress. <http://www2.ed.gov/about/overview/budget/budget16/justifications/index.html>.
- Wilson, W.J., 1987. *The Truly Disadvantaged: The Inner City, the Underclass, and Public Policy*. University of Chicago Press, Chicago.
- Wilson, W.J., 1997. *When Work Disappears: The World of the New Urban Poor*. Vintage.
- Wolpin, K.I., 2007. Ex ante policy evaluation, structural estimation, and model selection. *Am. Econ. Rev.* 97 (2), 48–52.

## CHAPTER 5

# Field Experiments in Developing Country Agriculture

A. de Janvry<sup>\*</sup>,<sup>1</sup> E. Sadoulet<sup>\*</sup>, T. Suri<sup>§,a</sup>

\*University of California, Berkeley, Berkeley, CA, United States

§MIT Sloan School of Management, Cambridge, MA, United States

<sup>1</sup>Corresponding author: E-mail: alain@berkeley.edu

## Contents

1. Introduction	428
2. A Review of FEs in Agriculture	429
3. Agriculture and FEs: A Conceptual Framework	437
4. Agriculture is Different: Implications for the Design and Implementation of FEs	441
4.1 Dependence on random weather realizations and risk	441
4.2 The spatial dimension of agriculture: heterogeneity and transaction costs	444
4.3 Seasonality and long lags	447
4.4 Market failures and nonseparability	450
4.5 Spillovers, externalities, and general equilibrium effects	452
4.6 Measurement	456
5. Discussion: Using FEs to Reveal the Production Function in Agriculture	460
References	463

## Abstract

This chapter provides a review of the role of field experiments (FEs) in answering research questions in agriculture that ultimately let us better understand how policy can improve productivity and farmer welfare in developing economies. We first review recent FEs in agriculture, highlighting the contributions they have already made to this area of research. We then outline areas where experiments can further fill existing gaps in our knowledge on agriculture and how future experiments can address the specific complexities in agriculture.

## Keywords

Agriculture; Developing economies; Field experiments

## JEL Codes

013; C93; Q1

<sup>a</sup> The authors are grateful to Shweta Bhogale, Erin Kelley, Gregory Lane, and Eleanor Wiseman for excellent research assistance, and to Mushfiq Mobarak, Abhijit Banerjee, and Esther Duflo for their reviews of the paper and suggestions for improvements. We also benefited from the review materials on FEs in agriculture prepared by Craig McIntosh, Rachel Glennerster, Christopher Udry, Ben Jaques-Leslie, and Ellie Porter. Errors and shortcomings are our own.

## 1. INTRODUCTION

Agriculture is an important sector in most developing economies, often the main form of employment for a majority of individuals and making up a substantial share of these economies' gross domestic product (GDP). For example, in 2014 for the low-income countries (as defined in the [World Bank \(2016\)](#) development indicators), agriculture made up 32% of GDP on average, declining to only 10% of GDP in 2014 in the middle-income countries. Similarly, in low-income countries, 70% of the population lives in rural areas, compared to 51% for middle-income countries. Agriculture may play a key role in the economic growth and industrialization of these economies and potentially have impacts on poverty reduction, food security, environmental sustainability, and community development ([World Bank, 2007](#)). While there are remarkable success stories of agriculture playing a key transitional role, these are few and far between. This raises an important question for research: what can be done to help low-income countries use agriculture to its full potential in achieving development?

A large research agenda to address this question has placed agriculture as one of the pillars of development economics, with chapters typically devoted to it in Handbooks of Development Economics and Agricultural Economics. However, many questions still remain unresolved due to the difficulties in establishing causality between determinants and outcomes in both positive and normative analyses. The use of field experiments (FE) offers an immense opportunity to address some of the outstanding issues in a rigorous and causal way. This chapter focuses on the role of FEs in this area, briefly describing contributions that have already been made using FEs, but also highlighting the specific complexities in the agricultural sector that such FEs have had to deal with and that leave room for further experimental research in this area. We discuss six specific features of agricultural environments and farmer behavior that have implications for the design of FEs and that illustrate the extensive room for further contributions to better understand the questions posed above. They are: (1) the dependence of outcomes on random weather realizations and exposure to risk, (2) the spatial dimension of agriculture with corresponding high heterogeneity and transaction costs, (3) the existence of seasonality and long lags in production, (4) the prevalence of market failures and their implications for farm household decision-making, (5) the occurrence of local spillover effects, externalities, and general equilibrium effects, and (6) difficulties in measurement. For each of these features, we describe in detail the issues in agriculture and then how FEs can be designed and implemented to close the remaining knowledge gaps in agriculture.

This chapter is structured as follows. In [Section 2](#), we review the FEs that have been used in agriculture in developing countries to provide a summary of the areas that experiments have contributed to as well as to highlight the findings from these studies. In [Section 3](#), we provide a broad conceptual framework to better understand

the role that FEs can play in agriculture. In [Section 4](#), we discuss each of the six areas where specific features of agriculture and the environment (both natural as well as economic) that farmers face have implications for the design and implementation of FEs. In [Section 5](#), we conclude by summarizing potential new areas of research on agriculture that may benefit from the use of such experiments, including using FEs to reveal the production function in agriculture.

## 2. A REVIEW OF FEs IN AGRICULTURE

Over the last decade, there has been a rapid growth in the use of FE to study agricultural issues in development. Supporting the more general growth of FEs in social science research, there has been a commitment from donors to better understand some of the outstanding questions in agriculture using this approach. A good example is the Agricultural Technology Adoption Initiative (ATAI) at the Abdul Latif Jameel Poverty Action Lab (J-PAL) and the Center for Effective Global Action (CEGA) that funds studies on the adoption and impact of technologies in agriculture. In the process to starting ATAI, a white paper was prepared on the then outstanding issues with technology adoption ([Jack, 2011](#)). Five years later, ATAI alone has funded 42 different studies in Africa and South Asia, across a wide range of topic areas.

The overwhelming majority of these studies have focused on the determinants of the modernization of agriculture and, in particular, the adoption, diffusion, and impact of technological and institutional innovations in agriculture. We summarize the experiments in agriculture by looking at the following topics: (1) the role of information about technologies in decision-making and how it affects farmer behavior, (2) the role of liquidity constraints and access to credit in technology adoption, (3) the availability of financial products and technologies to reduce farmer exposure to risks and mitigate the impact of shocks on farmers welfare, (4) the estimation of price responses and the role of subsidies in agriculture, (5) the role of access to price information for farmers, (6) contracts in agricultural value chains, and (7) the heterogeneity of conditions and types in influencing outcomes.

- (1) When does facilitating access to **information** about agricultural technologies to farmers make a contribution to agricultural development, and what is the most cost-effective way of doing this? Extension services or training are likely to be the most useful when new knowledge, not directly accessible to farmers, is needed to adopt a technological or institutional innovation. [Glennerster and Suri \(2015\)](#) showed that training was instrumental in the profitability of the New Rice for Africa (NERICA), a new rice variety with specific cultivation practices. Farmers who received both NERICA and training increased yields by 16%, while those with NERICA and no training experienced a small decline in yields. Information provided by extension agents can, however, be misleading to farmers if these agents

pursue a different objective function than that of farmers (typically maximizing yields instead of profit or utility) or do not properly account for the opportunity cost of farmers' labor time (for example, in recommending the adoption of the highly labor intensive System of Rice Intensification or the use of Conservation Agriculture [CA] without chemical herbicides). [Duflo et al. \(2008\)](#) thus found that properly timed and dosed fertilizers can be profitable in Kenya but that they were not profitable if used according to the dosages recommended by the Ministry of Agriculture. Similarly, farmers may fail to use available information, thus leaving money on the table. For example, studying seaweed farmers in Indonesia, [Hanna et al. \(2014\)](#) showed that farmers may be far from the efficiency frontier simply because they "fail to notice" some important aspects of the information they possess, such as the role of pod size as opposed to simply pod spacing on seaweeds.

Information can also be conveyed through social networks ([Foster and Rosenzweig, 1995](#)), though social learning may be limited by heterogeneity of farmer conditions and types. [Beaman et al. \(2014\)](#) analyze the factors that favor the adoption of pit planting in Malawi, a water harvesting technique commonly used in West Africa but largely unknown in Southern Africa. They found that social networks are useful to convey information but that they are the most effective when information is confirmed by more than one source. This suggests the need for several "seed" farmers for effective diffusion when introducing information on an innovation in a social network. [Cai et al. \(2015\)](#) randomized the training about a new weather insurance product in China. They found that social networks are effective at transmitting knowledge acquired by those who participated in intensive training sessions but not at sharing information about adoption decisions taken by particular individuals that is kept private. This suggests that information about individual decisions on the adoption of subsidized innovations, which is quite helpful for others in deciding to adopt, needs to be provided through public postings. FEs have thus been particularly effective at identifying the role of information in how farmers learn, both directly from what they do and indirectly from what others are doing. They have also shown that learning can be hampered by providing information to farmers which is not adapted to their own circumstances and objectives.

The point of entry matters. For example, [Emerick \(2014\)](#) finds that farmer-to-farmer diffusion of seeds in Odisha is constrained by the deep fragmentation of social relations in village communities along caste positions. A socially neutral door-to-door offering at market price secures a much higher level of uptake (40%) than what can be achieved through farmer-to-farmer diffusion (8%). [Ben Yishay and Mobarak \(2015\)](#) analyze the persuasiveness of different agents in communicating information on new technologies in Malawi. In this study, the

alternatives are government-employed extension workers, “lead farmers” who are educated and able to sustain experimentation costs, and “peer farmers” who are more representative of the general population and whose experiences may be more analogous to the average recipient farmer’s own conditions. Farmers find communicators who face agricultural conditions and constraints most comparable to themselves to be the most persuasive. [Duflo et al., \(2016\)](#) find little diffusion amongst the control group but that treatment farmers are more likely to adopt the more treated friends they have, highlighting a compounding or re-emphasis effect of the treatment.

Behavior in decision-making can be assisted through information provided under the form of nudges and reminders that countervail frequent behavioral traits such as time inconsistency, narrow bracketing in risk management, and failure to notice ([Hanna et al., 2014](#)). In the case of recommended fertilizer applications, the profit maximizing doses may be too complex or too knife-edge for farmers to take advantage of and so additional tools may be needed by farmers. For example, [Schilbach et al. \(2015\)](#) show that calibrated blue spoons can be used to simplify memorization of prescribed doses for farmers. [Casaburi et al. \(2014b\)](#) show that text messages can be used to remind sugar cane farmers when to use fertilizers in this case, treatment farmers had a 12% higher yield than control farmers. Simplifying decision-making and prompting behavior at the right time through simple reminders can thus be quite effective in optimizing the use of technological innovations. FEs have shown that farmers have a great degree of agency in decision-making on the use of their own resources, but at the same time, behavior can be quite complex, easily deviating from presumed rationality principles. There is therefore a unique role for FEs in helping reveal what motivates farmers in doing what they do.

- (2) **Credit** markets typically fail differentially more for poor people due to a lack of collateral. For this reason, much attention has been given to developing institutional innovations that can give farmers access to credit without the need for collateral. This has been the essence of the “microfinance revolution” ([Armendáriz and Morduch, 2005](#)), but this revolution has been particularly incomplete for agriculture where the turnover of loans is long and outcomes extremely risky. A surprising result, however, is that credit is often not the primary constraint to the adoption and profitability of innovations in agriculture as compared to risk reduction. For example, [Karlan et al. \(2014\)](#) use a FE that compares index-based weather insurance (subsidized or actuarially fair priced) with cash grants in Ghana. They find that, once insured, farmers are able to find the necessary financial liquidity to increase input expenditures (perhaps in part because lenders are more willing to extend loans when outcomes are insured for weather shocks), resulting in larger investments and the planting of more risky crops. Similar effects are not there for cash grants. The authors thus conclude that

risk was the binding constraint on farmers' investments, not a lack of liquidity. An FE with risk-reducing flood-tolerant rice technology showed a similar result: the risk reduction crowded-in additional investments and increased the use of credit from existing sources ([Emerick et al., 2016](#)).

A study conducted by [Ashraf et al. \(2009\)](#) to help farmers produce high-value export crops in Kenya found that lack of access to credit was not the main reason why farmers did not produce export crops and that those who were producing export crops had found access to credit on their own. Experimentation with post-harvest fertilizer purchases by [Duflo et al. \(2011\)](#) showed that behavioral incentives drove additional demand for credit rather than credit constraints limiting demand. Finally, the take-up of inputs tends to be low even when subsidized. [Carter et al. \(2014\)](#) found that only 50% of farmers took-up fertilizer vouchers in Mozambique.

[Jack et al. \(2015\)](#) study an in-kind asset collateralized credit product for water tanks for dairy farmers in Kenya and show take-up rates of about 42% when the loan requirements are minimal. These are high take-up rates for a credit product, considering the low take-up rates of microfinance loans ([Banerjee et al., 2015](#)). However, the study highlights the need for credit products and services that are tailored to the needs and particularities of farmers. Dairy farmers have much more regular incomes (they sell milk every day) than crop farming. Formal financial services such as banks largely stay away from agricultural credit in the developing world so that there is much room for innovation in the types of credit products available to farmers that account for the peculiarities of their income streams. More flexible lines of credit with repayment linked to the seasonality of sales can reduce costs and improve repayment rates. For example, [Matsumoto et al. \(2013\)](#) found a large demand for credit to purchase modern inputs for maize production in Uganda when credit repayment could be deferred until after the harvest. In [Beaman et al. \(2015\)](#), loan repayment was scheduled for after harvest, leading to a large demand by self-selected more productive farmers, and increased investment in inputs. Access to postharvest credit, still rarely available, could also be effective in helping farmers avoid selling at low prices at harvest time to subsequently buy at high prices when they run out of grains for consumption ([Burke, 2014](#))—this has yet to be fully tested. FEs therefore already have and can further be useful in the area of credit by experimenting with the design of new financial products better customized to the idiosyncrasies of agriculture.

- (3) **Index-based weather insurance**, in spite of its attractiveness in potentially helping deliver insurance to large numbers of smallholder farmers, has met with considerable difficulty in finding effective demand when not heavily subsidized. FEs have confirmed that, when actually used, index-based insurance can be quite effective at helping farmers better manage risk, reducing costly self-insurance, and increasing expected incomes, but take-up remains low at actuarially fair prices. [Schickele \(2016\)](#)

summarizes the results from 10 different FEs in four different countries on weather insurance. As they report, discounts and financial literacy interventions increase take-up of insurance, but at market prices, demand is low, in the range of 6–18% across these studies. Only a subset of studies measured changes in behavior due to insurance and in these, farmers were more likely to plant riskier (and higher yielding) crops, more profitable crops or invest in more inputs.

For example, [Mobarak and Rosenzweig \(2013\)](#) show that insurance helps Indian cultivators switch to riskier, higher-yielding crops. While beneficial to farmers, this switch, however, destabilizes employment opportunities for farm workers, suggesting the need to extend insurance coverage to them as well. In Ghana, [Karlan et al. \(2014\)](#) found that index-based insurance induced farmers to invest more in agriculture (increase fertilizer use, land cultivated, and total farming expenditures) and to select more risky activities (increasing the share of land planted with maize and decreasing that planted with drought resistant crops). [Cole et al. \(2014\)](#) gave away free rainfall index insurance policies in Andhra Pradesh, and as a result, farmers shifted production from subsistence crops to cash crops that were more rainfall-sensitive.

Similar results are found in broader insurance products. For example, [Cai et al. \(2010\)](#) found that insurance for sows significantly increased farmers' tendency to raise pigs in southwestern China, where sow production is considered a risky production activity with large potential returns. [Cai \(2016\)](#) finds (in a natural experiment) that expanding yield insurance for tobacco in China increased production. In Mali, farmers offered insurance for cotton planted more cotton ([Elabed and Carter, 2015](#)).

Given this, researchers tried to understand whether interlinked products may offer more value to farmers. [Gine and Yang \(2009\)](#) studied a bundled credit and insurance product but the bundled product had an even lower take-up than the credit product alone (18% vs. 33%). In Ethiopia ([McIntosh et al., 2013](#)), take-up was lower for a credit-insurance interlinked product than for an insurance only contract. We have yet to understand why the take-up rates for interlinked products are lower.

Overall, for insurance, FEs have contributed greatly to our understanding of such products over the last several years given the large number of studies that find similar results across different countries and context. Though they can have large impacts on farmers, demand at market prices is low. This opens the question of how to better design index-based insurance products and use incentive schemes to support learning and induce further take-up. Alternatively, insurance products could be a part of government welfare programs.

A more recent area has been to study whether risk-reducing technology may be easier for farmers to adopt than index-based insurance because it better corresponds to what they do and may have no extra cost. This technology can be effective not only

to cope with shocks but also to induce behavioral responses that can crowd-in other innovations and create incentives to factor deepening through adjustments in risk management. [Emerick et al. \(2016\)](#) found that flood-tolerant rice in India crowded-in the use of fertilizer and labor-intensive planting techniques. This downside risk-reducing technology thus has a double benefit: it reduces yield losses in bad years and increases yields in normal years. In this particular case, with one flood year expected every four, over the long run gains from increased investments in normal years (due to behavioral changes) were larger than gains from avoided losses in bad years (principally agronomic). FEs were here useful in revealing the full value of investing in agricultural research, with a lot more to be done to better understand such products in different contexts.

- (4) Setting price subsidies optimally or predicting the extent of effective demand for an innovation under market prices requires estimation of a full **price response** function. This is difficult to do due to the classical identification problem, i.e., the endogeneity of prices and simultaneity between supply and demand. FEs can be uniquely useful for this purpose. Recent results from experiments in other contexts have shown that poor people's demand for products beneficial to them tends to be highly price elastic around a zero price, falling rapidly to low levels often before the price reaches market equilibrium level ([Dupas, 2014](#)).

In agriculture, [Glennerster and Suri \(2015\)](#) showed the full price response to the uptake of NERICA rice in Sierra Leone, using varying subsidy levels, with take-up rates of 98% when the seed is free and rates falling to 20%, when the seed is offered at market price. In Ghana ([Karlan et al., 2014](#)) and in India ([Mobarak and Rosenzweig, 2013](#)), while a 75% subsidy rate can raise the take-up for index insurance to 60–70%, it falls to 10–20% at market price. [Cai et al. \(2016\)](#) found that Chinese farmers' demand for insurance is highly price elastic, achieving a 100% take-up when free but falling to 40% with prices still only equal to 70% of the fair price. This creates a huge problem in achieving market-driven take-up by poor people for potentially privately useful innovations. Inducing demand through short-term subsidies can create more opportunities to learn by increasing the likelihood of observing insurance payouts to oneself and to others in one's social network. [Carter et al. \(2014\)](#) found that subsidies in Mozambique not only induce short-term take-up but that demand persists in the long-run due to both direct and social learning. With subsidies highly costly and the design of optimum subsidies to induce take-up an important open question, FEs can be uniquely effective in estimating demand and experimenting with alternative subsidy schemes to match the complexity of learning, especially when outcomes are stochastic.

- (5) Providing **price information** to farmers has been hypothesized to be useful as farmers are typically poorly informed about market prices, particularly when transactions occur at the farm gate, far removed from markets. The question is whether

farmers will be able to use this information to obtain better prices when they sell their harvests, which is often not the case. Visaria et al. (2015) provided daily information via SMS to potato farmers in West Bengal on prices on local wholesale markets where traders sell the crops that they buy from them. They found that the provision of information did not affect the traders' average margins that ranged from 34% to 89%. Farmers altered the volume of their sales based on the price information they received, selling more potatoes when prices are high, and less when low. However, the finding of no impact on the traders' margins suggests that farmers have no direct access to wholesale markets and were thus not able to benefit from the information on prices to bargain with traders for better terms. Fafchamps and Minten (2012) similarly found no effect of market information delivered to Indian farmers through their mobile phones on prices received and cultivation decisions. In other cases, marketing incentives can be useful to change behavior or to solve information asymmetries about quality in markets for traders or sellers (not necessarily for farmers). For example, an FE in China showed that an innovation that improved a sellers' reputation by allowing consumers to recognize high-quality products through credible labeling induced them to differentiate higher quality watermelons, leading to higher profits and welfare for traders (Bai, 2015). Results from FEs thus show that access to price information largely does not make a difference in farmers production and sales decisions unless farmers have the ability to adjust where, when, and what they sell in response to that information, which is rarely the case. However, better information may be able to improve trader welfare. For farmers, instead, it seems that it is important to better understand the market structure of the traders that purchase from them—an open area where FE can yet contribute to our knowledge.

- (6) **Contracts** along value chains can induce smallholder farmers to switch to the production of high-value crops. Ashraf et al. (2009) show that contracts induced smallholder farmers to engage in the production of high-value crops in Kenya. Contracts may, however, be exposed to holdup behavior if not enforceable, and consequently be not sustainable as in the case studied by the authors. FEs can be used to explore the design of contracts that help reduce risk. In Kenya, insurance contracts were bundled in interlinked transactions between smallholder sugar cane producers and sugar mills. In this case, insurance costs were paid ex-post as they were deducted from final product payments by the sugar mill, securing high effective demand, a rare achievement for index-based weather insurance (Casaburi and Willis, 2015). Casaburi and Reed (2014) used an FE in Sierra Leone to analyze the extent of price pass-through in cocoa value chains where transactions link prices paid and credit contracts. In this context, intermediaries are both buyers of produce and providers of credit service. They found that raising trader wholesale prices were not passed through to farmers but were translated in a large increase in the likelihood that traders provided

credit to farmers. Price and credit pass-through can thus be substitutes. Isolating the price observation from the credit contract would underestimate the passing through of benefits to farmers from rising prices. Here again, FEs have been useful in identifying causal channels in the way value chains work.

- (7) It is well recognized that there is considerable **heterogeneity** in the conditions faced by farmers and farmer types themselves, implying that innovations, programs, and policies need to be correspondingly differentiated and targeted. Relevant heterogeneity, however, goes beyond observables. FEs can be used to control for nonobservables in constructing a counterfactual, and also to identify the role of nonobservables on outcomes through induced self-selection. [Jack \(2013\)](#) showed that allocation of subsidized tree-planting contracts across farmers through bidding in an auction helped self-select better farmers. This resulted in higher tree survival over a three-year period than allocation through random assignment. The gains from targeting based on private information led to a 30% cost saving per surviving tree for the implementing agency. [Beaman et al. \(2015\)](#) used a FE in Mali to show that farmers who took loans from a microlender had a higher return to investment than those who did not borrow. They did this by showing that the returns to a randomly assigned cash grant to a sample of nonborrowers who had previously been given the option to borrow were lower than the returns obtained by another population not offered loans at all and hence not self-selected out of credit. Finally, [Duflo et al. \(2011\)](#) used a FE to reveal farmer types with respect to the timing of the decision to purchase fertilizer in Kenya. They induced self-selection to show that 69% of farmers in their study area were stochastically present-biased, procrastinating in the decision to purchase fertilizer when they had available liquidity after harvest with subsequent underuse of fertilizers at planting time due to liquidity constraints. This result suggests that small, time-limited subsidies can be effective in nudging these particular types of farmers to optimize behavior toward fertilizer purchases.

If the heterogeneity in conditions and farmer types is important in making differential use of available technological and institutional innovations, FEs are uniquely useful to reveal the role of heterogeneity in influencing outcomes. This in turn can help design technological and institutional innovations that are optimally customized for farmers.

Clearly, there has been a rapid increase in the use of FEs in agriculture such that they have helped improve our understanding of the determinants of the adoption and impact of technological and institutional innovations in agriculture. These studies have started closing the large gaps in our knowledge of farmer behavior and agriculture but given the nature and complexity of agriculture, there is still much room for FEs to add to our knowledge base. There are high payoffs to well-designed FE that identify not only impact but also the causal channels at play. In addition, further experimentation and FEs can help shed light on implication for policy of the considerable degree of heterogeneity in agriculture. Next, we explored how future FEs in agriculture could be

designed to answer outstanding questions, starting with a conceptual framework to set the dimensions of the problem, and then proceeding to explore specific suggestions for the design and implementation of FEs.

### 3. AGRICULTURE AND FEs: A CONCEPTUAL FRAMEWORK

Natural phenomena play a large role in agriculture. In many developing economies, agriculture is mainly rainfed, which implies that there is often only one long growing season per year. In areas of the developing world that are irrigated (principally in Asia), while there are frequently two or three growing seasons per year, the seasonal patterns of production are sharply defined as crops grown in different seasons are not the same and interactions between them are important. Different crops have different growing season lengths and draw on different soil nutrients and are affected by plant diseases in different ways. All of this leads farmers to decide on an overall production plan for the year. Within one annual production cycle, the farmer will typically produce multiple crops over multiple plots and seasons and frequently also attend to a herd of different animal species. He may also work off farm to supplement his farm income.

In traditional modeling, one thinks of agricultural production as an implicit relationship between a vector of outputs  $Q$ , a vector of inputs  $X$  chosen by the producer, management practices or technology  $\theta$  also chosen by the producer, and a number of fixed factors  $Z^q$ , with output affected by a given weather realization  $W$ :

$$f(Q, X, \theta, Z^q, W) = 0 \quad (1)$$

In a context of perfect markets for output, inputs, credit, and insurance, the behavior of the decision-making unit (the household in most cases) would consist of choosing inputs to maximize expected profits, subject to constraints and the production function in (Eq. 1) given above. We use expected profit as the farmer has to optimize over his expectations of future prices as well as of weather outcomes. This leads to a set of optimal inputs and technology or management choices; each of which is in itself a function of the fixed factors, input prices,  $p^x$ ; expected output prices,  $Ep^q$ ; and the distribution function of weather outcomes,  $g(W)$ ; in addition to possible constraints. Output, in turn, depends on all the same variables and the weather realization in a particular year. A key issue in agriculture, and especially when producers face new technology, is imperfect knowledge or uncertainty about the production function itself, meaning that farmers' decisions are based on a subjective production relationship  $\tilde{f}$  instead of the true  $f$ .

However, given the more common scenario of imperfect markets for inputs and/or outputs, the quasiabsence of credit and insurance markets for smallholder farmers, and large transactions costs in rural areas, the household will instead maximize its utility over consumption (that includes home-produced and purchased goods and leisure), given some specific household preferences, subject to an internal equilibrium for

nontraded goods, a time constraint for own family labor, and a budget constraint over traded goods. An important aspect of agricultural production is its variability, due to both predictable seasonality and annual stochastic weather realizations or other shocks. Consumption smoothing concerns are thus important in household utility, suggesting that even in a simplified model one needs to consider the constraints in transferring goods or cash across seasons as well as across years. This then implies that the optimal input choice depends not just on the production characteristics described above, but also on household preferences, call them  $Z^c$ , and on constraints and opportunities outside the agricultural sector. Another set of considerations in decision-making, particularly in Africa, is due to the structure of households where men and women each have their own sphere of decision-making in agriculture.

All of this leads to a decision-making process that, even if one ignores for now all interannual dynamic relationships, implies many interrelated decisions. In a highly stylized fashion that will help categorize the different angles that research on agriculture has taken, we can formalize the decision-maker problem as follows:

$$\begin{aligned} & \max_{C_s, X_s, \theta, Q_s} EU(C_s, Z^c) \\ \text{s.t. } & \tilde{f}(Q_s, X_s, \theta, Z^q, W) = 0 \\ & \text{Non-traded products equilibria} \\ & \text{Time and budget constraints} \end{aligned}$$

and given opportunity cost of household resources outside the agricultural sector and prices.

where  $C_s$ ,  $X_s$ ,  $Q_s$  are vectors of consumption, input, and production chosen for different seasons  $s$ , respectively.

Note that these optimal input choices will vary from year to year and from location to location. Expected prices and weather realizations up to the time of input choice, and the expectations of weather realizations for the rest of the season until the outputs are produced, all vary from year to year and from one location to another. Prices all have an explicit spatial dimension, depending on harvests that, in turn, partly depend on weather realizations. Similarly, weather distributions and their realizations are very local in nature, both for a given year as well as in a given location.

A FE will typically introduce a treatment  $T$  affecting any of the exogenous decision factors represented in our framework by the fixed factors (physical and human capital endowments, property rights), the set of available technology and information on returns to inputs and technology (information, extension services, etc.), the constraints (access to credit or insurance), or the prices (subsidies, payment for environmental services, and contractual arrangements). To the extent that inputs and outputs are jointly determined, this treatment will usually affect all input decisions taken after the intervention, and then all outputs. Hence, measuring the outcome of a specific intervention will measure first

the specific output expected to be directly affected by the intervention but often also the impact on a variety of other outcomes that may be linked on the input side (which may well be all inputs, given that family labor is an input that spills across all activities). The producer's response to any treatment that aims at inducing some increase in output could produce substitution into specialization at the cost of decreasing other activities. Or conversely, a strong shock on any input or output could have a very small effect if the farmer reoptimizes all his choices and spreads the shock over all activities.

The average treatment effect can be simply computed by the difference in means of the selected outcomes  $Y$ . However, given the integrated decision-making process, it is very likely that the many dimensions of heterogeneity described above will translate into heterogeneous treatment effects. As extreme examples, the promotion of a labor intensive technology may have no effect on households that are labor constrained; information on prices will not affect households that are too far removed from markets and hence have no bargaining power with traders who come to their farm gate; a drought resistant crop variety will have no benefit (and even potentially a yield penalty) in a year where there is normal rainfall.

The general expression for the conditional impact will thus be:

$$\begin{aligned} & E[Y_{ilt}|(T = 1), Z_i^q, Z_l^q, Z_i^c, Ep_l^q, p_l^q, p_l^x, g(W_l), W_{lt}] \\ & - E[Y_{ilt}|(T = 0), Z_i^q, Z_l^q, Z_i^c, Ep_l^q, p_l^q, p_l^x, g(W_l), W_{lt}] \end{aligned} \quad (2)$$

where  $Y_{ilt}$  is the outcome for farmer  $i$  in location  $l$  in year  $t$ ; the  $Z^q$  are fixed factors that either vary by individual ( $i$ ) or location ( $l$ ); the  $Z^c$  are household preferences;  $Ep_l^q$  are the farmer's expectations of output prices  $p_l^q$ ;  $p_l^x$  are the input prices;  $g(W_l)$  is the distribution of weather for that location, and  $W_{lt}$  is the weather realization for that location in year  $t$ .

There is a strong inter-relationship between the multiple inputs and outputs of any agricultural activity and a potential dependence of these decisions on consumption, which implies that any treatment could have broad impacts across multiple dimensions of a given household.

Finally, the spatial dispersion of agriculture and the presence of high transaction costs could create local economies, with concentration of economic activity within the localities and some isolation from the rest of the larger economy. This implies that there may be important spillovers, externalities, and general equilibrium effects. In the framework developed above, for example, an intervention facilitating the accumulation of stocks of harvested products will affect local prices  $p_l^q$  and  $p_l^x$ , which spread the benefits to all sellers in the community and reduce the benefit of arbitrage for the beneficiaries (Burke, 2014).

This presentation of the agricultural production process highlights key areas that FEs in agriculture could build on and contribute further to our knowledge base on: (a) the critical role of weather in affecting production and the treatment effects themselves, such that multiple realizations over time would be useful to better understand expected

impacts; (b) the spatial heterogeneity in physical and economic contexts, and the trade-offs between getting precise results on homogenous groups and testing for heterogeneous effects across these contexts; (c) the seasonality and long lags in the production process, that can impose high demands on the information collected by researchers but that also introduce opportunities for tracing out how shocks and changes may affect behavioral responses across seasons; (d) market failures and the nonseparability of household decisions that create an additional dimension of heterogeneity that is hard to characterize and that we still do not know enough about; (e) social spillovers, environmental externalities, and general equilibrium effects that allow us to better understand the broader impact of the intervention through tailored measurement of such effects; and (f) gaps in our knowledge of measurement that stem from the necessity of observing the quantities of so many inputs, outputs and their prices to fully measure impact and the channels of causality in even a fairly targeted and simple intervention.

What makes agriculture potentially unique is the level of agency farmers have over what they do, exacerbated by the degree of heterogeneity under which they operate. In other sectors such as education or health, beyond deciding which facilities to use, individuals rely on the school or the hospital to have the necessary knowledge for good decision-making. Parents choose the school their children enroll into but then largely delegate to the teachers the practice of education. In contrast, farmers tend to believe that they know how to farm and, even more, know what specifically works best for their idiosyncratic piece of land. They thus make multiple decisions about the problems they face such as choice of crops and cultivation methods; and they may be reticent to follow good advice from an extension agent or community members because they believe that it does not apply to their own circumstances. Whether farmers have the right or wrong knowledge, they are typically faced with more choices and are thus required to make more decisions than agents in other sectors. Bad decisions can easily be taken. Understanding how farmers decide based on their particular inner vision of the processes they live in is thus both uniquely important to the study of agriculture and what makes it such a fascinating topic to analyze.

Due to its integration with the lives of households, agriculture is clearly more than just a sector of economic activity. We saw above that in the context of imperfect markets, consumption and production decisions are strongly connected. An intervention encouraging specific cropping patterns may have as its main objective improving the family's nutrition and health. Other interventions may have the explicit purpose of reducing the labor burden on children or of affecting the balance of power between genders in the household. Because of its very large dependency on natural resources, agriculture is also strongly related to the environment. Interventions encouraging certain practices may be seeking to enhance the long-term sustainability of resource use, such as soil, water, and biodiversity conservation. The range of outcomes of interest thus spans a very large domain.

FEs can be very useful to precisely measure the impact of specific treatments and the channels involved on these multiple outcomes. Other topics are not so easily studied through FEs. Examples are the long-term effects of technological change (because of general equilibrium effects on prices) and agricultural policy interventions (because of lack of degrees of freedom). For this, other approaches are necessary, such as natural experiments capitalizing on the rollout of policies or discontinuities in treatment. Quite often, a natural experiment can be complemented by a FE, for instance to measure a particular impact on behavior or to experiment with the design of a complementary intervention.

## 4. AGRICULTURE IS DIFFERENT: IMPLICATIONS FOR THE DESIGN AND IMPLEMENTATION OF FEs

### 4.1 Dependence on random weather realizations and risk

As highlighted above, the outcomes of most interventions and farmers decisions in agriculture depend on the specific realizations of rainfall and more generally of weather, especially in rainfed agriculture. In sub-Saharan Africa, for instance, 93% of arable land is rainfed. Ignoring for now the other elements of heterogeneity in the general expression (2) given above, we should think of the outcome  $Y$  in an agricultural FE with treatment  $T$  and weather realization  $W$  as  $Y(T, W)$ . For any given weather realization,  $W_{lt}$ , which varies over both space and time, the conditional average treatment effect of the intervention is:

$$ATE(W_{lt}) = E[Y_{ilt}|T = 1, W_{lt}] - E[Y_{ilt}|T = 0, W_{lt}]$$

The treatment effect of interest is either the average treatment effect for a given location,  $ATE_l$ , which integrates  $ATE(W_{lt})$  over the intertemporal distribution of weather  $g_l(W_{lt})$  for that location; or the average treatment effect for the area represented by the selected locations, which is the double integral of  $ATE(W_{lt})$  over both space and time:

$$ATE = \iint ATE(W_{lt})g(W_{lt})dW_{lt}$$

where  $g(.)$  is the overall distribution of  $W$  over locations and time. With this simple notation, we can describe how the weather interacting with interventions and farmers' decisions materialize in FEs and are areas future FEs can build on. In particular, we highlight four important considerations.

First, in any particular year, the average treatment effect over the space of the experiment is given by the cross-sectional distribution of  $W_{lt}$  over locations,

$$ATE_t = \int ATE(W_{lt})g_t(W_{lt})dW_{lt}$$

where  $g_t$  is the distribution of weather over locations in year  $t$ . This may be of limited interest since it informs neither the meaningful  $ATE_l$  that may influence the uptake and actions of agents in location  $l$  nor the overall  $ATE$  of interest to the researcher or policy makers. There are, however, situations where computing  $ATE(W_{lt})$  is feasible and will be useful, when there is a wide range of random weather shocks over the sample area in this particular year. The precision of the conditional results will depend on the density of observations at each weather realization. Using secondary data on the time series of weather events (easily available) in specific locations allows the computation of  $ATE_l$  by integrating over the distribution of weather realizations in that location. An illustration of this are the results in [Dar et al. \(2013\)](#) and [Emerick et al. \(2016\)](#) on the impact of flood tolerance on rice yields across the number of days of flooding. Here, the conditional results were relatively easy to compute given the geographically dense cross-sectional variation in the occurrence and duration of flooding across farmers' plots.

Second, due to the length of the production cycle in agriculture, we typically only observe a few weather realizations over time for a particular location as part of a FE. This implies that we may not be able to compute well the conditional  $ATE_{lt}$  over the whole distribution of weather for each particular location. As a consequence, we may only learn about limited segments of the ATE function for particular climatic events that happen to have been observed. Third, weather is a multifaceted event that is difficult to characterize. This multidimensionality makes it harder to know what matters for a given weather realization in terms of affecting the observed outcome. For example, the variables used to measure the stress of low rainfall on rice production in India are potentially the date of the onset of the monsoon, the cumulative rainfall over different phases of the growing season, and the number of contiguous days without rainfall during the flowering period. We also know that temperature matters, as measured by degree-days, as well as a number of other factors such as wind speed and hours of sunlight.

Fourth, incomplete information about the impact of an intervention due to the role of weather events also has implications for how farmers understand the relation between a given intervention and its outcomes. Farmers' understanding of the value of a new intervention may be based on only one or a few weather realizations, making it difficult for them to get a precise estimate of the returns to the induced action. For example, [Beaman et al. \(2013\)](#) argue that "if the signal on the profitability of fertilizer is weak relative to the noise resulting from weather variability, it will be hard for farmers to learn about how much—if any—fertilizer is optimal for them to use on their particular plot of land given other possible constraints they face on inputs (including labor, for example)." Learning-by-doing may therefore be conditional on the realization of weather outcomes. Communicating with others about the outcome of the intervention may also be imperfect as the weather realization that conditioned the outcome is difficult to characterize. For FEs that give importance to behavioral responses, there are gains to documenting

what farmers may have been able to learn by their own doing and what they learn from others. In the particular case where the intervention is a risk-reducing technology or a weather insurance product, the timing of the weather event relative to the intervention may affect the inference that both the researcher and the farmer can draw from the intervention. A risk-reducing technology typically has a yield penalty in normal years. Similarly, an insurance product has a premium to be paid even in normal years. If a normal year occurs before a shock year, this penalty creates a negative wealth effect for subsequent years. A series of normal years may discourage adoption, altering observed behavior when a bad year occurs. And selection into treatment may keep in the farmers who are more risk averse and less liquidity constrained. By contrast, a bad year occurring before a good year may create a wealth effect such that understanding the mechanisms underlying the effect of the intervention will be important. In Emerick et al. (2016), the technology tested had no yield penalty. It so happened that a bad year occurred first, identifying the shock-coping value of the innovation. In this case, the first year wealth effect from a seed minikit was sufficiently small not to have meaningful effects on the second year behavioral response. The second year response that occurred under normal weather therefore allowed them to identify the pure risk management response to a reduction in downside risk.

As we think of designing future FEs in agriculture, the above particularities with regards to weather highlight how such experiments can be designed to make further contributions to our understanding of agriculture:

- (1) *Clustering:* Some researchers may want to cluster treatment and control observations to have as much as possible the same weather outcomes. For example, treatment and control can be located close to the same meteorological station from which the weather realization will be observed by the researcher. This, of course, needs to be done without compromising the risk of spillover effects between treatment and control that we discuss in more detail below.
- (2) *Dispersion:* To obtain several conditional ATEs will require spreading the experiment over wide geographic areas to observe different weather realizations. The more covariate weather events are, the more geographically dispersed the experiment should be. However, the trade-offs of spreading experiments over large areas include the potential heterogeneity in other factors across a wide area and the difficulty of managing the experiment and collecting data in very distant locations.
- (3) *Duration:* Having FEs that run across several time periods to get impacts under different realizations of shocks over time will add great value to our understanding of what works in agriculture, though other factors may interfere with observed outcomes. For example, we may need to allow for learning effects so as not to confound these with outcomes of the particular intervention under study. External conditions will also undoubtedly change, and they may be endogenously affected by the diffusion process itself. The trade-off here is potentially difficulty in maintaining a control group, especially if there is diffusion.

- (4) *Information sharing:* If there is rapid loss of learning when there is a sequence of events that makes the innovation unnecessary, then large experiments with information sharing across distant locations will be necessary to preserve the value of learning. This also implies that there will be a lot of fluctuations in individual behavior over time, with many individuals moving in and out of using a particular technology (similar to [Suri, 2011](#)), with implications for power calculations in experimental design.

## 4.2 The spatial dimension of agriculture: heterogeneity and transaction costs

Over the last decade, some of the research in agriculture has highlighted how important local conditions are in economic decision-making. Though that is now becoming well recognized, it has also opened up a number of avenues for future research where FEs can play a role.

This heterogeneity of local conditions and their impacts may be quite complex in agriculture for two reasons. First, there are biological or ecological (soil quality for crops) and environmental (climate and altitude) differences across space that make the impacts of farmers decisions (such as which technology to use, whether to use fertilizer, how much fertilizer to use, which crops to grow, and which livestock to raise) very heterogeneous ([Suri, 2011](#)). Second, there are the more standard dimensions of how the structure of the economy and economic policy affect (broadly) transaction costs. Effective producer prices differ widely according to location and market power and likely from year to year. Factor prices, such as for fertilizer, differ across farmers and may also differ from year to year, especially if these factors are being imported. The implication is that returns from a particular technology or investment may differ widely, making it attractive for adoption for some farmers but not for others, and more broadly affecting farmers' decisions and outcomes. Furthermore, as described above, poorly developed irrigation and farmers dependence on rainfall as their primary source of water implies that not only do soil and climate conditions matter at any given point in time, but they are continually changing and changing differentially across space which has impacts on farmers decisions ([Suri, 2011](#)).

This heterogeneity raises complexities not just for the farmers themselves but also for the extension agents that advise them. First, an understanding of the true returns of any technology when those returns depend on local environmental and soil conditions is difficult. This is illustrated in the low level of tailoring in recommendations provided to farmers by scientists and extension services (see, for example, [Duflo et al., 2008](#)). So, there is a need to develop extensive agricultural R&D systems to build up the knowledge base of agricultural technologies (the production function), and this is an area of potential collaboration between social scientists and hard scientists, which we discuss in [Section 5](#). Second, the focus on yields is a crucial first step, but ultimately it only paints

half the picture for the farmer. What matters for the farmer is profits, which are harder to measure well, as we discuss more below. Any given new technology or investment could (and if it increases yields should) increase labor costs and may well involve other costs (such as costs of other complementary inputs, effects on soil quality as the technology may draw different nutrients from the soil, etc.). The costs of accessing the technology might be quite different across space due to variation in linking to markets, infrastructure or, more generally, transaction costs in both input and output markets.

Our understanding of how these external factors affect returns to technologies and investments is still not complete and is therefore an important opportunity for FEs. They can be designed to measure conditional impacts, where conditionality is in terms of the dimensions of heterogeneity that matter for the outcomes of interest. Knowledge of conditional impacts will perhaps also allow better customization of recommendations to the specific conditions under which a particular farmer operates, potentially achieving larger benefits from the recommendations than can be achieved through recommendations based on average effects.

As above, if  $X$  measures the dimensions of heterogeneity that matter for the outcome  $Y$  of interest, conditional impact is measured as  $ATE_X = E(Y|T = 1, X) - E(Y|T = 0, X)$  and the  $ATE$  is given by:

$$ATE = \int_{X \in Expert-mental Pop} ATE_X f(X) dX$$

This particular  $ATE$ , though useful, has some limitations: (1) it is not informative for all individuals within the population and, in some cases, may even not apply closely to any one in particular if there is considerable heterogeneity in the population; (2) it may be measured with a large standard error; and (3) it may not be able to be used out of sample.

The potential for FEs is immense here, as they can be designed to study which dimensions of heterogeneity matter, describing them with selected indicators, and characterizing their distribution in the population. FEs have an important role to play in gathering the wealth of information needed to make policy prescriptions in agriculture as useful as possible, allowing recommendations to be customized to population sub-groups, approximating precision farming, with large potential payoffs for farmers.

When using FE to add to the knowledge base in this area, the following design issues may be important:

- (1) *Heterogeneity of spatial conditions:* Ideally, future FEs would cover a large heterogeneity of spatial conditions so that researchers can better understand whether the benefits of the given farmer decision or investment under study vary by soil type or microclimate. However, this will come with costs, budgetary as well as the time needed

to design, manage, and study an intervention that extends across a large geographical area. An alternative is to work within homogeneous subpopulations, in a way where the homogeneity can be measured based on observable characteristics, leading to findings that can be generalized over these dimensions. Examples of such an approach include [Carter et al. \(2014\)](#) who chose to only work with farmers located along major roads, [Burke \(2014\)](#) who only worked with the highly selected clients of his partner organization, the One Acre Fund and [Glennerster and Suri \(2015\)](#) who worked with upland rice farmers.

- (2) *Characterizing heterogeneity:* There has been little work on (and there is therefore a lot of room for) measuring some of these external factors that matter. An example would be to try to measure soil quality differences to get a sense of local heterogeneity and how much that matters. We know little about how to measure soil quality well, but there are surely many lessons to be learnt from agronomists. There are few studies actively measuring soil quality, with the exception of [Fabregas et al. \(2015\)](#) who study soil testing and the demand for it among farmers in Western Kenya. An alternative is to measure an observable outcome that predicts soil quality deficiencies, as done by [Islam \(2014\)](#), for example, who uses leaf charts in Bangladesh that show how well the plant is taking in nutrients from the soil. In addition, how soil conditions vary over time with climatic events, and with the past use of fertilizers, and whether this affects treatment outcomes is an open question. On the climate side, rainfall data are more easily available, both across space and over time. These data can be incorporated to better understand the role of microclimates in agriculture.
- (3) *Heterogeneity along unobservable dimensions:* Producers are also heterogeneous along unobserved dimensions (such as an intrinsic productivity) and in the effort that they are willing to apply to a new process and hence may enjoy very different benefits from getting access to new technologies. This may explain either large noncompliance or observed low average impact of adoption. In reality, one may be interested in measuring the impact of access to a technology for those who would eventually adopt it rather than for a random sample of the population. Experimental designs can be geared to reveal such heterogeneity and/or foster the desired selection of participants to the experiment. The theory of two-step randomizations to reveal and measure unobserved heterogeneity is developed by [Chassang et al. \(2012\)](#) who suggest the use of “selective trials” as a generalization of RCTs. These selective trials include a mechanism that let the agents reveal their own valuation of the proposed technology and may furthermore incentivize effort in order to disentangle the role of the product itself from that of the effort applied to its use. [Chassang et al. \(2013\)](#), for example, invited farmers to bid to enhance their chances of winning a new mechanical technology, allowing them to measure the impact of the technology on those most eager to try it. Future FEs can generalize such approaches,

ensuring that the selection mechanism reveals the unobservable characteristics of researcher interest and not other constraints such as wealth or access to liquidity.

- (4) *Heterogeneity and learning:* Similar heterogeneity may exist along the dimensions that explain the diffusion of new technologies. If individuals are very different from each other (or their plots are of very different qualities), then there may be limited room for learning. For example, [Conley and Udry \(2010\)](#) found that farmers are more likely to learn from those who are more similar to them or who have a similar experience. Similarly, [Tjernström \(2014\)](#) finds that soil quality heterogeneity (at the village level) makes farmers less likely to respond to their peers' experiences of a new technology. There is still much room to design experiments to better understand the structure of technology diffusion in agriculture, especially with regards to the degree of similarity with specific others in social networks, and how the degree of analogy matters for learning.

### 4.3 Seasonality and long lags

The seasonality of agriculture for cereal or staple crops production makes economic decisions in agriculture more complicated than for, say, a high-turnover vegetable producer. Farmers make investments before or during their planting season, with no returns (i.e., no incomes) until months later when the crop is harvested. Because agricultural cycles depend on rainfall, there is also a commonality in the cycle across all farmers—they all tend to plant and to harvest around the same time. This implies that their demands tend to be quite correlated, especially with regards to labor. Their sales and purchases of food products are also correlated, with a sharp cycle of prices on local markets if there is imperfect tradability. This means that the farmer's realizations and expectations of output prices are not just indexed by location and year, but also vary within a given year depending on when a farmer chooses to sell. This long cycle in agricultural production also allows for partial revelations of uncertainties and adjustment of behavior along the year. Once rains get started, farmers will update their decisions about the optimal time to plant ([Kala, 2014](#)) and what inputs to use. As more of the weather realization is revealed, the farmer will continue to adapt his input choices accordingly.

More specifically, these seasonal cycles pose a number of issues for farmers. First, farmers face long delays between expenses and revenues. This makes access to credit both all the more important as well as all the more difficult. All the more important because farmers often have to make investments that are costly such as fertilizer purchases long before they earn incomes. All the more difficult to manage as standard microfinance loans with regular payments every week or month are not relevant since households may not have a regular income flow. Second, farmers are often engaged in multiple activities spread over time, so that labor calendars are an important factor in their decisions. For example, if land needs to be cleared for planting, the demand for labor at planting will

be high and will involve family labor, possibly supplemented by hired labor. The same is true for harvesting and processing of the crops—both family and hired labor are used. However, in between these two periods, the main activities are crop management, i.e., activities like weeding that are often largely conducted by household labor, partly because less labor is required and partly because monitoring is important for such activities (Bharadwaj, 2010). Designing labor calendars to smooth demands on family labor time becomes an important issue in crop and technological choices (Fafchamps, 1993). This also implies that there will be associated strong seasonalities in other relevant economic variables like wage rates. Third, due to the long time lags between actions and outcomes, there may be time inconsistencies in decision-making that can have impacts on productivity, sales, and incomes. Commitment devices may be important to overcome these inconsistencies. In this perspective, Duflo et al. (2011) explore a savings commitment device for fertilizer purchases, Brune et al. (2016) a commitment savings product with crop sales paid directly into a bank account as opposed to cash, and Casaburi and Willis (2015) an insurance scheme in Kenya with premiums paid ex-post after harvest.

Fourth, the timing of the agricultural cycle leads to an accompanying sales and nutrition cycle in a lot of developing economies. There seems to often be a market failure in crop storage in that most of the farmers who sell grain will tend to sell it at the time of harvests. This implies that sale quantities are high and prices low at harvest. Over the subsequent months, the price of food rises so that right before harvest, there is less food available, and its price is quite high. There is evidence that this leads to seasonality in consumption—a lot of farmers report having a hungry season before harvests come in where food availability and consumption are lower than at other times of the year. This seasonality in food consumption could also lead to seasonality in nutrition outcomes of farmers and their household members though this has not been well or widely documented in the literature. Some new agricultural technologies try to shorten the growing cycle of the crop with the hope of allowing multiple cropping over the year with a shorter hungry season, e.g., NERICA rice seed (see Glennerster and Suri, 2015). It remains an open area of research to understand how farmers should best smooth seasonal fluctuations. One option has been to test ways of improving access to storage (for example, Casaburi et al., 2014a,b; Burke, 2014). However, storage itself may only be part of the story. Similarly, the hungry season may force farmers into second-best activities such as seasonal migration and participation to the labor market (Fink et al., 2014) that may impact on or take away from the productive activities in agriculture, particularly when technology is labor intensive.

On the research side, this seasonality has been important to account for. First, there is a wide heterogeneity in crops and their seasons. Most farmers will invest in a multitude of different crops, with decisions made about one crop affecting others. A technological change in any of these crops (such as adoption of a short duration variety) may affect decision-making in all other crops. Although there may be a season for a lot of cereal and

staple crops like maize, wheat, and rice, there are other crops which do not follow such straightforward cycles and may still be a very important part of farmers' livelihoods. Tree crops like cocoa and coffee are an example, where the tree needs approximately five years to grow before it is productive and so investments are made years before there is a return to them. Another example is cassava which once planted can be left in the ground for multiple years if needs be—farmers often use this as a backup food crop when harvests of their main food crop are poor. Second, it implies long lags between the implementation of any intervention and the outcomes, extending the time horizon for a research project. This implies that experiments in agriculture will often span multiple years and are therefore riskier in nature (for example, the second round of the Glennerster and Suri NERICA experiment has been on hold due to the Ebola outbreak in Sierra Leone) with longer-term outcomes harder to measure.

Given this seasonality and time-lags, future FEs could build on the following areas:

- (1) *Understanding seasonality*: Monitoring outcomes from a given intervention on a seasonal basis rather than an annual basis could provide useful learnings, filling in important steps in the process, and reduce dependence on recall data that may be noisy ([Beegle et al., 2012](#)). For example, any one intervention may impose short-term seasonal costs, with potential longer-term gains or vice versa. Measuring all these is important for welfare interpretations. Adoption of a labor-intensive technology may reduce seasonal migration, with a seasonal cost that is compensated by a higher annual gain. Similarly, data on inputs and behavior should be collected during key periods of the agricultural process. An example is the work of [Goldstein and Udry \(2008\)](#) in Ghana who organized data collection with a round of surveying every six weeks. It is important to note the role that new technologies can play in facilitating this sort of higher frequency data collection: the wide adoption of cell phones in developing economies makes it possible to collect data on labor use, financial transactions, and product sales and purchases by frequent calls. There may also be interesting uses of sensors that give real-time data, for example, on moisture conditions and on labor movements. We discuss this in more detail in the section below on measurement.
- (2) *Successive adjustments*: FEs that focus on interventions early in the season could be designed to allow adjustments or complementarities later in the season. This would greatly enhance our understanding of complementarities and substitutions across seasons, an area where little is known to date.
- (3) *Cost of delays*: Studies in agriculture have built in the time lags inherent in agriculture. Missing the planting season by just a few weeks for any reason can have the heavy cost of a full year lost in the research process. As the timing of planting varies from year to year, for any intervention that needs to happen before planting (like introduction of a new seed or fertilizer), researchers have had to build in enough time to allow for a potentially early planting season that given year. The

flip side is true for situations when planting happens later than expected in a given year. Finally, a single year may not be enough time to see effects, as discussed above.

#### 4.4 Market failures and nonseparability

In the developing country context, production and consumption decisions are integrated. Households cultivate small areas of land and usually provide most of the labor needed on their farms. These households maximize utility, subject to production, and so may not have a single objective defined on their production outlays, such as expected profit or a function of the distribution of net profit to take into account risk and seasonal patterns of income. A household is said to be nonseparable when its decisions regarding production (adoption of a technology, use of inputs, choice of activities, and desired production levels) are affected by its consumer characteristics (consumption preferences including over leisure, demographic composition, etc.).<sup>1</sup> Many papers have demonstrated that a separability test typically fails for smallholder farmers across the developing world (for example, [Benjamin, 1992](#); [Jacoby, 1993](#)). The conceptual framework above [see [Eq. \(2\)](#)] highlighted how any treatment  $T$  introduced to a farmer can affect not just input decisions and hence, outputs, but that the response will also depend on household preferences if separability fails.

Some potential symptoms of nonseparability are households producing much of what they eat, using exclusively family labor for certain tasks, having excess family labor that cannot find outside employment, etc. These correspond to situations where the household is constrained in the amount of labor or food that it can exchange on the market, or where transaction costs on markets are sufficiently high that the internal equilibrium shadow price of the commodity or factor produced and used by the household makes it suboptimal to either purchase or sell it ([Renkow et al., 2004](#)). For example, [Fafchamps \(1993\)](#) shows that farmers select an overall annual cropping pattern that maximally smoothens family labor needs throughout the year. Similarly, farmers maintain the cultivation of several varieties of rice with different maturation lengths or planted at different times to spread the harvest time ([Glennerster and Suri, 2015](#)). Farmers often choose to maintain the production of their main staple food (including in India where they have access to cheap government subsidized rice), limiting availability of resources for potentially more profitable cash crops.

Since the behavioral models of separable and nonseparable households are quite different, we can expect heterogeneity in the uptake of and response to certain agricultural interventions along this dimension. Whether this is important or not of course

<sup>1</sup> They leave aside the cases where risk attitude and time discount are the only “preference” elements that enter the production function, since these can easily be included in a production model.

depends on the intervention. For example, the take up of a new technology may be low because farmers cannot sell any excess output they produce or acquire the labor they need for a labor intensive technology (see Beaman et al., 2013). Alternatively, a farmer may not be able to respond optimally to changes in the price of crops because they are constrained to produce what they want to consume and to use labor as available in the family, which might prevent the household from reallocating its land toward more profitable crops (de Janvry et al., 1991).

There are two relevant aspects of nonseparability. First, it is not directly or easily observable, as opposed to say distance to a city or soil quality or rainfall. Whether a household is separable or not is the result of an equilibrium that depends on its resources and preferences. One can sometimes use proxies for this. For example, if the main constraint is on the labor market side (either on sale or purchase), the ratio of available family labor to farm size may provide some indication of the likelihood that the household may be separable or not (Beaman et al., 2013, show that high return to agriculture in Mali is associated with large household size). Future work in this area would account for the quality of the labor force and alternative occupations that determine the opportunity cost of working on the farm and the fact that being “nonseparable” is partially an endogenous choice of the household.

Second, separable and nonseparable households may differ not only in terms of the intensity of their technology uptake or price response but potentially also across the channels that explain the impact of any intervention. Key with separability is transaction costs in markets that make the household into a net seller or a net buyer of a particular food item or of labor. Key with nonseparability is resource endowments and internal demand that affect the shadow price of this resource or product. As long as the intervention affects the shadow price without making it affect the effective farm-gate market price as seller or buyer, the household will remain in the nonseparability status (Sadoulet and de Janvry, 1995).

Future FEs in agriculture can be designed to shed more light on the issues around separability in the following ways:

- (1) *Household versus farm as the unit of analysis.* Studying the household as a whole as the unit of analysis will add great value. While any intervention in agriculture may be targeted at the production process of a given crop, hence just the farm operations of a household and perhaps even a given plot, studying the effect on other household decisions can add to what we know about household separability and hence be important in assessing overall outcomes or welfare and shedding light on potential future dynamics around market access.
- (2) *Data on production and consumption.* There is yet a lot to learn about how interventions will affect both the production and consumption decisions of households. To this end, surveys around a FE could collect detailed information not only about farmers’ production but also about their consumption decisions. This can help establish how

an intervention targeted at production might be affected by the farmer's consumption preferences. Similarly, an intervention targeted at consumption, such as a guaranteed employment scheme or a food subsidies program, will have potentially important effects on production decisions.

- (3) *Characterizing household-specific market failures.* One other open gap in this area is characterization of prevailing market failures which may well vary from household to household. Particular interventions, for example, that reduce transaction costs and link households to value chains, may transform households from the nonseparable to the separable status, with potentially large implications for technology adoption, the elasticity of supply response, and ultimately farmer incomes.
- (4) *Using FE to test for separation.* Testing for separability relies on detecting a causal effect of household consumption characteristics on production behavior, and it has until now been conducted using instrumental variable techniques. An innovative use of FEs could be to reveal the nonseparability of household behavior by, for example, looking at response to an intervention that affects consumption patterns, such as the effective price of food or the opportunity cost of family labor. These would only affect production decisions if the household were nonseparable.

## 4.5 Spillovers, externalities, and general equilibrium effects

In rural settings, people are closely linked to each other, implying that interventions on some individuals may generate a wide range of spillover effects, externalities, and possibly general equilibrium (GE) effects. While externalities are not a specificity of agriculture (and hence will be found in many other chapters), spillover effects in the diffusion of new technologies, externalities on the environment, and general equilibrium effects are particularly important in agriculture. In the context of agriculture, we can classify these community links as personal ties among farmers or their households, cost transfers among neighbors through the environment, or price changes through general equilibrium effects in local markets. We discuss each of these in turn briefly.

First, a treated farmer will likely inform others in his social network that he is cultivating something new, changing his farming techniques, or adopting a new insurance scheme. This will impact others' behavior in the community through learning, imitation effects, and potentially economies of scale. Second, there may be large local environmental externalities in agriculture. For example, the effectiveness of biological control methods depends on whether the fields around yours are similarly managed; the cost of underground water depends on the intensity of pumping of farmers extracting water from the same aquifer as you. Finally, as farmers go to sell their products on local markets, hire labor, and earn some additional income in off-farm activities, prices and wages may change more broadly. Looking back at Eq. (2) above, for example, externalities and GE effects can affect all aspects of farmers' decisions via their impacts on the cost side, via input prices (such as wages) and expected output prices.

Though these externalities and GE effects may contradict the stable unit treatment value assumption, they should be seen as interesting in themselves, and can, in fact, be a key part of the experimental design. We separate this discussion into three separate areas. First, we look at spillover effects from social interactions, in particular, farmer learning and the diffusion of agricultural technologies. In rural settings, individuals are keenly aware of what their neighbors are doing because they can physically observe their actions. In addition, agricultural communities are closely knit social groups. Individuals within these communities share strong social bonds, which can take the form of informal mutual insurance arrangements, joint production, product sharing, or altruistic behavior. As a result of these two factors, there will undoubtedly be direct spillover effects from one treated farmer to other farmers in the community. The learning (and diffusion) externality around new agricultural technologies has been extensively studied, highlighting when and how learning effects may be important (see the literature review section above). For example, we know that the point of entry matters for diffusion ([Emerick, 2014](#); [Ben Yishay and Mobarak, 2015](#); [Beaman et al., 2014](#)). Similarly, the complexity or riskiness of the new technology matters. Some practices or technologies may be easier to learn than others. For example, [Tjernström \(2014\)](#) shows that the extent to which social networks can be relied upon to transmit information depends on the quality of the information environment in which individuals operate. [Glennerster and Suri \(2015\)](#) on the other hand find little evidence of diffusion of a new rice seed in Sierra Leone. Although the literature in this area is vast as highlighted above, there is still room to learn about the extent to which diffusion and learning spillovers will depend on the specific structure of the community and its surroundings. Ultimately, there is not overwhelming evidence showing farmer-to-farmer diffusion which leaves large gaps in our understanding of how farmers may learn and an area where FEs could continue to make important contributions.

Second are environmental externalities. There is an additional set of externalities that arise in agriculture that come from crop diseases, pests, water, soil conservation investments, etc. Decisions farmers make about any investments that change the use of water and that affect crop diseases and pests can have broader impacts than just on their own farms. Though there are many examples of such practices (such as integrated pest management, water extraction, etc.), there have been few FEs to study these spillovers. Water is a very specific resource in agriculture, partly because of its importance to enhance agricultural productivity and partly because it is a communal resource. There are therefore large externalities imposed by water extraction, both across space as well as on future generations (see [Foster and Sekhri, 2007](#), for an example). In addition, the short-run and the long-run benefits of improved water access are unclear because it is a communal resource across generations (see [Hornbeck and Keskin, 2015](#)). There are clearly large gaps in the literature here where FEs can be designed to help better understand the magnitude of these externalities and their implications for agricultural investments.

Third, we discuss the role of market level and GE effects. Agricultural markets are generally small and can be relatively isolated from larger trading networks due to high transaction costs. Local farmers produce a few main crops and sell any surplus they have in nearby markets. Labor markets are typically local and land markets even more so. In addition, given the externalities described above, FEs may be designed as cluster RCTs, i.e., where the level of randomization is a village or a community rather than an individual. This implies that general equilibrium effects inside the village may well be pervasive and may also occur via nonagricultural outcomes. If the farmers treated by any given intervention see a rise in their agricultural income, they will spend more money on other services provided by the community. This will increase revenue flowing to other nonagricultural activities and contribute to market level changes in prices. Evidence for GE effects is given by [Jayachandran \(2006\)](#) who shows that productivity shocks (in particular, rainfall shocks) cause large changes in the district level wage in India. Similarly, [Mobarak and Rosenzweig \(2014\)](#) document the GE effects of rainfall insurance on agricultural wages. The flip side of these isolated markets is that they reveal that transaction costs between markets are high, and that there is immense potential for arbitrage, but that there are market failures (potentially credit markets for traders or traders being imperfectly competitive) that restrict these arbitrages from happening. Any given intervention that affects the trading or market environment may well have impacts on prices, and hence broader general equilibrium effects.

Understanding these effects is important when wanting to predict the impacts these interventions would have at scale. To date, just a few studies have tried to measure these effects, with much room in the literature for more. Examples include [Burke \(2014\)](#) who tries to measure the price impacts of a crop storage intervention and on-going experiments by [de Janvry et al. \(2015\)](#) on the labor market effects of drought resistant rice seeds in Jharkhand. In a non-FE setting, [Evenson and Gollin \(2003\)](#) showed that prices decreased a lot due to the Green Revolution, and this benefited consumers. In general, farmers only benefited from this if there were significant reductions in production costs that were greater than the fall in prices.

Finally, there may be strong feedback loops that result in GE effects. The short-term effects of many interventions will differ from their long-run impacts. A priori it may be difficult for research to predict what will happen over longer time horizons. As the agricultural and economic systems adapt to a new intervention, many of the original gains from the intervention may be mitigated. It is important for policy makers to be aware of these effects before they invest in an intervention strategy that they hope will have long-run benefits. This scenario is conceptualized in the famous theory of the “technological treadmill” ([Cochrane, 1993](#)): early adopters might see larger profits, as they are the first to have access to the new technology. As more and more farmers gain access to the new technology, and “get on the treadmill,” the initial adopters will see their profits decrease. For example, [Burke \(2014\)](#) finds positive effects of a storage intervention in

areas where there was a low density of treatment. However, in areas with higher treatment density, these positive effects were significantly reduced due to general equilibrium price changes. As prices decline on local markets, the short-run early adopters' gains typically measured in RCTs will tend to be rapidly dissipated and transferred to consumers on insulated local markets via lower prices, with an ultimately very different incidence of benefits from technological innovations.

All these issues have implications for the design of future FEs. Considering first the case of externalities or spillovers:

- (1) *Designing the FE to avoid or measure spillover effects.* FEs should be designed to capture the spillovers that arise (from either learning externalities or the more natural crop disease/pest/water externalities described above) or work to design controls that are largely unaffected by these externalities. On the former, careful designs can be implemented to accurately measure spillovers (e.g., [Baird et al., 2014](#)). The latter would likely affect the level of randomization, which will have cost and power implications.
- (2) *Informing spillovers.* The FE could collect information to measure the effects on surrounding members within the community (even if the unit of analysis is at the village level). This involves measuring effects on farmers that are not treated by the intervention. For example, [Glennerster and Suri \(2015\)](#) designed a FE to measure spillovers in the NERICA technology, partly because of learning but also because farmers could just share the harvest of the improved rice with their neighbors.

Considering the case of general equilibrium effects and feedback loops:

- (1) *Powering FEs to measure GE effects.* Measuring general equilibrium effects is challenging. A lot of interventions are small, and so, the effects they are likely to have on market level prices or other GE outcomes are small, which poses power problems. Future work could focus on larger interventions or designs that vary the concentration of treatment across (say) villages to measure these effects. In addition, some of the effects on nonagriculture could be measured directly through collecting broader data.
- (2) *Using market surveys to inform GE effects.* Some of these effects will operate through markets. FEs could incorporate the design of market level surveys. These surveys can record prices in nearby markets, as well as wage levels. Researchers could try to measure welfare effects on consumers via falling prices in local markets. As prices fall in local markets, some net seller farmers who did not adopt the technological innovation may start entering the category of more subsistence farmers.
- (3) *Measuring the cost side of GE effects.* The full set of cost reductions that happen due to any agricultural technology changes may be hard to measure. Researchers conducting FEs in this area could collect detailed data on as broad a possible set of cost reductions due to a technological intervention as possible.

- (4) *Using demand-side interventions to preserve early adopter gains.* With shallow local markets, the short-term and long-term benefits for adopters of any yield increasing or cost-reducing innovation can be dramatically different. This stresses the importance of linking more effectively local to global markets for the benefits from technological change to be at least partially retained by farmers. An open area for new FEs is therefore to study how supply-side interventions could be complemented with demand-side interventions that may effectively deepen local markets ([McIntosh, 2014](#)).
- (5) *Sustaining and scaling-up FEs to measure GE effects.* FEs could extend over longer periods of time and be scaled up over larger numbers of adopters to trace out how the profitability of the intervention changes as more and more farmers adopt. They can therefore measure whether the benefits to new technologies wear out, and how often new technologies need to be brought in to keep the technological treadmill going.

## 4.6 Measurement

A FE will typically aim at measuring how an intervention affects an agricultural outcome along a number of dimensions. As mentioned above, inputs and outputs in agriculture interact in complex ways. For example, promoting the production of a particular cash crop may affect the production of all other crops, promoting fertilizer use on a specific crop may affect the use of other inputs on this crop and also potentially the whole production process. It is therefore important to measure the full set of multiple inputs and outputs, in itself a daunting task due to the extent of the information that needs to be collected and the difficulty of properly characterizing some of these inputs and outputs. In addition, observing the margins on which the farmer has adjusted to the intervention is important in itself, as it can highlight the channels through which the intervention led to the aggregate reduced form effect. For example, [Emerick et al. \(2016\)](#) show how the shock-coping gains from a new risk-reducing rice variety generate further benefits through behavioral responses in risk management. [Beaman et al. \(2013\)](#) show a reduction of the marginal effect of one input through adjustment on another margin.

One challenge is to aggregate these outcomes into a performance indicator. In the tradition of agricultural economics, this is done by considering restricted profit that measures the return to the fixed factors or the value of all outputs less than the cost of all variable inputs, i.e., using the notation from above,  $\Pi = pQ - qX$ , which involves: (1) establishing the distinction between fixed and variable factors, (2) measuring all variable inputs, and (3) measuring all prices that are needed for the aggregation. Finally, while a properly designed FE will balance fixed factors over the different treatment arms to allow measuring an average treatment effect over the distribution of these factors, there may be some first-order heterogeneity in the impact of the intervention that ought to be considered so that it may be important to measure these fixed factors.

In what follows, we suggest some areas that can help improve the design of future experiments and enhance future learnings:

- (1) *Concepts difficult to inform.* In many cases, farmers may not know the quantities of inputs used or the quantities of outputs produced. This is particularly the case for an output that is at least partially consumed and that is harvested over a period of time. A good example is cassava that stays in the ground throughout the year (and potentially even multiple years) and is only harvested when needed for consumption, or milk that is collected daily or even twice a day. In this case, surveys cannot ask about total production over the last 12 months as farmers do not think of their production of cassava or milk in this way ([Carletto et al., 2015](#); [Zezza et al., 2014](#)). In other cases, the process may be so complex that farmers are not even aware of whether what they do is important (for example, [Hanna, et al., 2014](#)). This implies that researchers may have to directly observe input use and/or production rather than ask farmers for it. For the case of continuous use, the researcher may need to rely on recalls or else opt for more frequent data collection (see [Goldstein and Udry, 1999](#)). The pervasiveness of cell phones in the developing world will make it much easier to conduct high-frequency phone surveys (see [de Janvry et al., 2015](#)), though this is not without its own problems as frequent interviews may result in survey fatigue and attrition.
- (2) *Quantifying self-provided inputs.* Many inputs are self-provided by the household, for example, household labor. If the household was completely dedicated to agriculture, household labor could be considered as a fixed factor, but this is generally not the case. An additional complication for labor is in defining the proper unit in which to measure labor use, because the work is irregular and dispersed. A labor-day is not a concept used by farmers themselves, and hence, they do not necessarily know how to quantify it. Similarly, detailed accounting of time use has proved very difficult to do (see [Jack et al., 2015](#) on how they collected time use and their pilots for various ways of collecting time use). Perhaps, new electronic tracking devices will provide opportunities to improve data collection on time use.
- (3) *Measuring prices.* Prices may be observable but often have strong seasonal variations. In a world of well-functioning markets, neither the spatial nor the seasonal variation of prices presents a conceptual measurement problem. Spatial variation would reflect heterogeneity in transaction costs and seasonal variation the cost of storage. When there are market failures, with strong seasonal variations, the price chosen to value output can make a difference on the presumed profitability of a farmer. [Duflo et al. \(2008\)](#) chose to price maize at the level it reaches just before the next season's harvest, as most farmers are then net maize buyers and purchase maize at the end of the season after their own stocks have run out. [Beaman et al. \(2013\)](#) chose to value output at producer prices at the time of harvest. They argue that this, "avoids

confounding a potential increase in profits from increased output, with the returns to storage". [Burke \(2014\)](#) tracks the full-time path of maize prices between two harvests since he is interested in the potential returns from arbitraging through storage, with farmers going from sellers when prices are low to buyers when prices are high. Alternatively, [Falcao et al. \(2014\)](#) use traders to obtain high-frequency prices from local markets. This is not only an issue for output prices but also for input prices. In general, there is a deficit of information on product and factor prices. If the researcher collects these as self-reports from the farmers, they need to be careful about what set of prices they ask farmers about and whether these are the relevant prices for the problem at hand.

- (4) *Pricing family labor.* When family labor is sold on the casual labor market, then the opportunity cost of work on the household's own farm is the wage on that casual market at that particular time (i.e., work on the farm effectively competes with labor allocation outside the farm; [Jack, 2013](#)). However, even in this best-case scenario where there is participation by family labor in the casual labor market, particular on-farm activities may not compete with the off-farm time allocation. In this case, the true cost of family labor is the shadow price on their time. If family labor does not work at all off the household farm, then it can be considered a fixed factor shared by all household activities. The shadow price in each activity (and in agriculture) is the equilibrium price internal to the household and is hence endogenous and not directly observable. It is well known that a large number of family farms do not seem economically viable when family labor is valued at the observed market wage rate in the casual labor market, implying that this is not the correct way to value family labor. However, valuing it at the correct shadow price is extremely difficult. In FEs, we are often more interested in the direction of change in profitability due to a particular intervention than in the absolute value of profits. One solution would therefore be to use a range of labor prices from market price to a fraction of that price and measure the corresponding range of changes in restricted profits as a sufficient indicator of impact. That said, there is room for research that is directed at better understanding agricultural labor markets and the role and productivity of family vs. casual labor in these settings.
- (5) *Valuing livestock.* Many outputs may be difficult to define and/or measure, especially livestock. Livestock plays an important function as a store of value, a producer of organic fertilizers for crops, as well as being a source of income (milk and meat). For many small farmers, their livestock forms a major component of their asset base. Yet, because of the variety of species and ages, the variation in animal quality and herd dynamic, it may be difficult to measure the asset value of the livestock herd. In addition, how their value evolves over time is complex as it depends on investments made by the household as well as age effects on productivity. There has been a lot less research on livestock, which implies that a lot of the measurement

issues for livestock have yet to be experimented with and may be more complex than for crops.

- (6) *Measuring soil and seed quality.* Returns to agricultural technologies and inputs are highly dependent on soil quality, which is often completely unobserved to the researcher. The returns varying with soil quality will, in turn, affect the demand for the new technology. As the use of FEs in agriculture grows, measuring both true soil quality as well as farmers' perceptions of it may help advance our understanding of returns. Surveys can elicit farmers' perceptions but these may give an incomplete picture of the underlying soil properties. More recently, studies have taken soil samples to help characterize soil quality and potentially discover additional properties of the land not known to the farmer (see [Fabregas et al., 2015](#); [Mahajan et al., 2014](#)). There is still a lot to learn about the relevant variation in soil fertility across plots and over time and how this relates to the past use of fertilizers and recent weather events. The same applies to the genetic content of the seeds in use. DNA testing can be used to reveal the origins of seeds, but these tests are expensive, and the extent of relevant heterogeneity in genetic content is not known. These two aspects of soil and seed quality could make a large difference in the customization of plot-level recommendations.
- (7) *Observing technological change.* Many FEs in agriculture study the adoption, diffusion, and impact of technological change. There is much interest in this question for public investment and donor accountability in investing in agricultural research. However, observing technological change can be complex. New seeds may come under the form of a rapid succession of new releases, with each new vintage only making a marginal improvement over the previous. As a sequence, there may be large gains over time, but each release only makes a marginal contribution. This makes it unclear what the appropriate counterfactual is. Only every so often do we observe truly transformative technologies, such as IR6 rice and semidwarf wheat cultivars that underpinned the Green Revolution. When "nature does make jumps," measuring the return to investment in research would use the original traditional varieties as a counterfactual, varieties that may disappear quickly. Future FEs could be geared to better understanding the dynamics of successive adoption.
- (8) *Are double-blind trials useful in agriculture?* A recent measurement issue that has been raised is whether the equivalent of double-blind trials is useful for FE in agriculture. Researchers have begun to think about behavioral responses to an intervention or a technology in agriculture as separate from the true yield or genetic returns to the technology (see [Bulte et al., 2014](#), for example). This distinction between genetics and behavior is, however, at best, murky. A lot of technologies are only beneficial when they are accompanied by certain practices. These practices typically involve a change in farmer behavior, and they are part of the technological package. In addition, any technology that increases yields for a farmer will automatically change some

aspects of farmer behavior, for example, labor since there will be more crops to be harvested, and therefore, more labor will be required. This is a change in behavior that should be bundled with the technology. The distinction between a behavioral response on the part of the farmer and the technology impact is clearer in the case of [Emerick et al. \(2016\)](#) where they find that the technology helps reduce the yield losses due to flooding (including some element of farmer behavior as recommended by the extension agent). In addition, farmers respond the following year to this change in the risk profile of their outcomes which implies that the second year behavioral response is a lower bound to the role of behavior in contributing to the yield gain of the new technology. In general, there is likely little scope for blind tests in agricultural FEs.

- (9) *The plot as a unit of analysis.* Finally, keeping track of plot-level panels is a challenge as plots are not a static concept. The characterization of production with homogenous conditions is typically at the plot level. Flooding for instance is very much a plot-level event. If the farm is fragmented, as typical in smallholder farming where land has been inherited and divided across family members over generations, a particular household will own several plots with very different features, the very reason why there has been fragmentation ([Foster, 2014](#)). Working with plot-level panels would therefore be ideal. However, the definition of a plot is a fluid concept that changes over time. Farmers may not necessarily know how much fertilizer and other inputs they have applied to each plot, while they know how much they have purchased overall. Production from various plots may be combined for threshing, making identification of output per plot difficult. A farmer may change the crop mix on a plot over time, invalidating the previous year's definition of a given plot and making it impossible to match plots over time, even if one had good GPS maps of plots (which is costly and time-consuming).

## **5. DISCUSSION: USING FEs TO REVEAL THE PRODUCTION FUNCTION IN AGRICULTURE**

As illustrated above, agriculture is in many aspects different from other productive sectors. Unlike the case of manufacturing, where the production function is a construct of the enterprise, we do not know enough about the true structure of the production function in agriculture. And, because of its link to nature, the production function for agriculture may have more in common with the production function for health than with a standard manufacturing process. One can marvel at the fact that a teaspoon of productive soil typically contains more than one million species of living organisms that are interacting in producing soil fertility. It is not always clear what the full list of inputs into the production function for agricultural products may be—land, labor, climate, and soil quality are a small subset of the elements that could possibly matter

and each of these cannot easily be represented by a single-scalar measure. As an example, the timing of input use is very important for the ultimate outcome. This implies that the combination of quantity and timing of inputs already gives us a long list of entries. We also need to know more about how these inputs may interact with each other in the production function for output: there are almost surely complementarities between labor and soil quality, and between labor and climate. What exact form these complementarities take is not well understood. In the past, the production function has been specified as the product of two subfunctions: a classical production function with the contribution made by factors of production, and a damage function, where damage is done by factors such as pests and bad weather ([Lichtenberg and Zilberman, 1986](#)), though this is likely incomplete.

In addition, we know little about what farmers themselves know about this production function or how they approximate what this function may be when making their decisions. They may have a very limited or error laden estimate of what the production function looks like, which implies that they are continuously learning themselves and making decisions based on incomplete information. And yet, as [Schultz \(1964\)](#) taught us, knowing the production function is essential to assess the marginal product of factors and make optimum decisions on factor use.

There are three levels at which FE research can help reveal the production function. The first is Agricultural Experiment Station (AES) research. Agronomic research has been a pioneer in using statistical experimentation, typically randomization in a Latin Square design analyzed in an analysis of variance framework. It was famously introduced into agronomic research by R.A. Fischer at the Rothamsted Experimental Station in Great Britain. Greco-Latin Square designs can be used to test two-by-two combinations of experimental treatments, such as seeds and fertilizer doses. A high-yielding variety seed and a traditional seed can thus be tested against various levels of fertilizer use. The problems with AES experiments are twofold. First, the conditions under which the experiment is conducted are generally not reported. For example, what were the levels of irrigation water and pesticides applied in the seed-fertilizer experiment? This limits the applicability of the results obtained. In this case, the particular segment of the production function that has been revealed is not clearly specified. Second, expectedly, with yield as the reported agronomic outcome, the conditions under which the experiment is conducted (such as water, pesticides, and labor practices) are for maximum yield. This does not correspond to what farmers will subsequently do in their farms, where their objective functions will be profit maximization or utility maximization if there is risk aversion. Comparing AES experimental yields with yields in farmers' fields not surprisingly tend to find that the former are larger than the latter.

The second level of useful analysis is in demonstration plots set up by extension agents in farmers' fields. The farmer is coached by the agent in applying a number of production practices to a new technology. The rest of the production decisions are

left to the farmer. The advantage of this approach is that it brings the technology close to potential adopters, with their own objective functions and production conditions. The disadvantage is that the demonstration plot may not always be accompanied by a counterfactual technology (Hancock, 1992). Farmers may be left to compare the treatment outcome to what they do on their own farms, which is for each farmer the next best technology that typically differs from farmer to farmer. There may also not be learning in demonstrating the response function of the new technology. There is room to use the demonstration plot approach to identify the production function by bringing this approach closer to that at AESs. Farmer selection could be formalized, and every demonstration plot farmer could be requested to define a counterfactual technology and cultivate it in an adjacent plot. Information on impact measured by difference could then be diffused across farmers so that there is an opportunity to learn about the response function.

The third is FEs in farmers' fields through randomized control trials. This is the typical seed minikits approach that Dar et al. (2013) followed in testing the flood tolerance value of SwarnaSub1 rice. Minikits are distributed randomly in treatment villages, farmers plant these seeds in a plot of land of their choice (though ideally this should be a randomly chosen plot), they apply their own self-selected cultivation practices in accordance with their objective function, particular climatic events occur, and a yield or any other outcome is observed. If there is a cross-sectional range of climatic events (such as days of flood duration), we can observe the yield advantage of Sub1 over the farmer's counterfactual seed for that range of events. The advantage is that the outcome corresponds to what farmers will be doing with the technology for their own purpose. The disadvantage is that we only learn about an average treatment effect and some heterogeneity, not the full production function. Because it is difficult to characterize the conditions and events under which the measurement was made, it may be difficult for the farmer to learn from the observed outcomes, and it is difficult to use the information to help others learn from these farmers' outcomes.

We see two innovative roles that FEs can play in research on the production function in agriculture. First, experiments can be designed to reveal the production function. We can gradually uncover the input responses, ideally over a wide range of differing environmental and climatic conditions and begin to better understand what the agricultural production functions may look like. This will of course require an almost heroic effort and may lack some of the glamour that motivates economists. A second role that RCTs can play which may be equally important in the short run is to better explore and understand what the implicit production function is that farmers are using in making their own decisions. Here, agronomy meets the social sciences. Understanding this would give us a window into better understanding farmer epistemology, learning and behavior, farmer constraints, and how best can policy respond to inform agents, alter behavior, and relax constraints.

FEs can be a powerful tool for project design and policy recommendations in agriculture. The specificities of the production process in agriculture and the agency and social relations in which it is embedded imply an emphasis on design and measurement in these FEs, matching survey data to data on climactic and soil conditions, so that we can ultimately address questions of first-order importance. We highlighted opportunities for areas where future FEs may be able to contribute a great deal to the existing stock of knowledge in this area. Over time, there have been and, no doubt, there will continue to be immense improvements in design and measurement as we adapt to the learnings of earlier experiments and studies. This will be accompanied by improvements in collecting and matching data, including using innovative tools for measurement. As described, FEs have already contributed important results that help our understanding of the role that agriculture plays in development. The remaining gaps in how to use field experiments in agriculture to address important outstanding development issues create a promising future research agenda.

## REFERENCES

- Armendáriz, B., Morduch, J., 2005. *The Economics of Microfinance*. MIT Press Books, Cambridge, MA.
- Ashraf, N., Giné, X., Karlan, D., 2009. Finding missing markets (and a disturbing epilogue): evidence from an export crop adoption and marketing intervention in Kenya. *Am. J. Agric. Econ.* 91 (4), 973–990.
- Bai, J., 2015. Melons as Lemons: Asymmetric Information, Consumer Learning, and Seller Reputation. Working Paper, Department of Economics, Massachusetts Institute of Technology.
- Baird, S., Bohren, A., McIntosh, C., Özler, B., 2014. Designing Experiments to Measure Spillover and Threshold Effects. Policy Research Working Paper Series 6824. The World Bank.
- Banerjee, A., Karlan, D., Zinman, J., 2015. Six randomized evaluations of microcredit: introduction and further steps. *Am. Econ. J. Appl. Econ.* 7 (1), 1–21.
- Beaman, L., Ben Yishay, A., Fatch, P., Magruder, J., Mobarak, M., 2014. “Making Networks Work for Policy: Evidence from Agricultural Technology Adoption in Malawi.” Working Paper. Northwestern University.
- Beaman, L., Karlan, D., Thuysbaert, B., Udry, C., 2013. Profitability of fertilizer: experimental evidence from female rice farmers in Mali. *Am. Econ. Rev.* 103 (3), 381–386.
- Beaman, L., Karlan, D., Thuysbaert, B., Udry, C., 2015. “Selection into Credit Markets: Evidence from Agriculture in Mali.” Working Paper. Yale University.
- Beegle, K., Carletto, C., Himelein, K., 2012. Reliability of recall in agricultural data. *J. Dev. Econ.* 98 (1), 34–41.
- Ben Yishay, A., Mobarak, M., 2015. Social learning and incentives for experimentation and communication. *Rev. Econ. Stud. Forthcoming Review of Economic Studies*.
- Benjamin, D., 1992. Household composition, labor markets, and labor demand: testing for separation in agricultural household models. *Econometrica* 60 (2), 287–322.
- Bharadwaj, P., 2010. Fertility and rural labor market inefficiencies – evidence from India. *J. Dev. Econ.* 115, 217–232.
- Brune, L., Giné, X., Goldberg, J., Yang, D., 2016. Facilitation savings for agriculture: field experimental evidence from Malawi. *Econ. Dev. Cult. Change* 64 (2), 187–220.
- Bulte, E., Beekman, G., Di Falco, S., Pan, L., Hella, J., 2014. Behavioral responses and the impact of new agricultural technologies: evidence from a double-blind field experiment in Tanzania. *Am. J. Agric. Econ.* 96 (3), 813–830.
- Burke, M., 2014. “Selling Low and Buying High: An Arbitrage Puzzle in Kenyan Villages.” Working Paper. University of California at Berkeley.

- Cai, H., Chen, Y., Feng, H., Zhou, Li-A., 2010. Microinsurance, Trust and Economic Development: Evidence from a Randomized Natural Field Experiment. NBER Working Paper No. 15396.
- Cai, J., de Janvry, A., Sadoulet, E., 2015. Social networks and the decision to insure. *Am. Econ. J. Appl. Econ.* 7 (2), 81–108.
- Cai, J., de Janvry, A., Sadoulet, E., 2016. Subsidy Policies with Learning from Stochastic Experiences (Working Paper).
- Cai, J., 2016. The impact of insurance provision on households' production and financial decisions. *Am. Econ. J. Econ. Policy* 8 (2), 44–88.
- Carletto, G., Jolliffe, D., Banerjee, R., 2015. From tragedy to renaissance: improving agricultural data for better policies. *J. Dev. Stud.* 51 (2), 133–148.
- Carter, M., Laajaj, R., Yang, D., 2014. Subsidies and the Persistence of Technology Adoption: Field Experimental Evidence from Mozambique. NBER Working Paper No. 20465.
- Casaburi, L., Glennerster, R., Suri, T., Kamara, S., 2014a. Providing Collateral and Improving Product Market Access for Smallholder Farmers: A Randomised Evaluation of Inventory Credit in Sierra Leone, 3ie Impact Evaluation Report 14, July 2014.
- Casaburi, L., Kremer, M., Mullainathan, S., Ramrattan, R., 2014b. Harnessing ICT to Increase Agricultural Production: Evidence from Kenya. Harvard University.
- Casaburi, L., Reed, T., 2014. Interlinked Transactions and Pass-through: Experimental Evidence from Sierra Leone. Stanford Institute for Economic Policy Research, Stanford University.
- Casaburi, L., Willis, J., 2015. Time Vs. State in Insurance: Experimental Evidence from Contract Farming in Kenya. Working Paper. Stanford University.
- Chassang, S., Miquel, G.P.I., Snowberg, E., 2012. Selective trials: a principal-agent approach to randomized controlled experiments. *Am. Econ. Rev.* 102 (4), 1279–1309.
- Chassang, S., Dupas, P., Snowberg, E., 2013. Selective Trials for Agricultural Technology Evaluation and Adoption: A Pilot. ATAI Pilot Project. CEGA, University of California Berkeley.
- Cochrane, W., 1993. The Development of American Agriculture: A Historical Analysis. University of Minnesota Press, Minneapolis, MN.
- Cole, S., Giné, X., Vickery, J., 2014. How does risk management influence production decisions? Evidence from a field experiment. *Rev. Financial Stud.* (Forthcoming).
- Conley, T., Udry, C., 2010. Learning about a new technology: pineapple in Ghana. *Am. Econ. Rev.* 100 (1), 35–69.
- Dar, M., de Janvry, A., Emerick, K., Raitzer, D., Sadoulet, E., November 22, 2013. Flood-tolerant rice reduces yield variability and raises expected yield, differentially benefitting socially disadvantaged groups. *Sci. Rep.* 3. Article number 3315.
- de Janvry, A., Fafchamps, M., Sadoulet, E., 1991. Peasant household behavior with missing markets: some paradoxes explained. *Econ. J.* 101 (409), 1400–1417.
- de Janvry, A., Sadoulet, E., Emerick, K., Dar, M., 2015. The Impact of Drought-tolerant Rice on Local Labor Markets in India (Work in Progress).
- Duflo, E., Kremer, M., Robinson, J., 2008. How high are rates of return to fertilizer? evidence from field experiments in Kenya. *Am. Econ. Rev. Pap. Proc.* 98 (2), 482–488.
- Duflo, E., Robinson, J., Kremer, M., 2011. Nudging farmers to use fertilizer: theory and experimental evidence from Kenya. *Am. Econ. Rev.* 101 (6), 2350–2390.
- Duflo, E., Keniston, D., Suri, T., 2016. Technology Adoption in Agriculture: Evidence from a Randomized Experiment in Rwanda (Work in Progress).
- Dupas, P., 2014. Getting essential health products to their end users: subsidize, but how much? *Science* 345 (6202), 1279–1281.
- Elabed, G., Carter, M., June 2015. Ex-ante Impacts of Agricultural Insurance: Evidence from a Field Experiment in Mali. Working paper.
- Emerick, K., de Janvry, A., Sadoulet, E., Dar, M., 2016. Technological innovations, downside risk, and the modernization of agriculture. *Am. Econ. Rev.* 106 (6), 1537–1561.
- Emerick, K., 2014. The Efficiency of Trading in Social Networks: Experimental Measures from India. Working Paper. Tufts University.

- Evenson, R., Gollin, D., 2003. Assessing the impact of the green revolution, 1960–2000. *Science* 300 (758), 1078710.
- Fabregas, R., Kremer, M., Odendo, M., Robinson, J., Schillbach, F., 2015. Evaluating Agricultural Information Creation and Dissemination in Western Kenya (Work in Progress).
- Fafchamps, M., 1993. Sequential labor decisions under uncertainty: an estimable household model of West-African farmers. *Econometrica* 61 (5), 1173–1197.
- Fafchamps, M., Minten, B., 2012. Impact of SMS-based agricultural information on Indian farmers. *World Bank Econ. Rev.* 26 (3), 383–414.
- Falcao, L., Gertler, P., McIntosh, C., 2014. E-Warehousing for Smallholder Farmers. ATAI Project. JPAL.
- Fink, G., Jack, K., Masiye, F., 2014. Seasonal Credit Constraints and Agricultural Labor Supply: Evidence from Zambia. NBER Working Paper 20218.
- Foster, A., Rosenzweig, M., 1995. Learning by doing and learning from others: human capital and technical change in agriculture. *J. Political Econ.* 103 (6), 1176–1209.
- Foster, A., Sekhri, S., 2007. Can Expansion of Markets for Groundwater Decelerate the Depletion of Groundwater Resource in Rural India? (Brown University Working Paper).
- Foster, A., 2014. Spillovers, Coordination Failures and Land Fragmentation. Working Paper. Brown University.
- Giné, X., Yang, D., 2009. Insurance, credit, and technology adoption: field experimental evidence from Malawi. *J. Dev. Econ.* 89 (2009), 1–11.
- Glennerster, R., Suri, T., 2015. “Measuring the Effects of NERICA, Short Duration Rice, on Harvest Prices.” ATAI Project. MIT.
- Goldstein, M., Udry, C., 1999. Agricultural Innovation and Resource Management in Ghana. International Food Policy Research Institute, Washington, DC.
- Goldstein, M., Udry, C., 2008. The profits of power: land rights and agricultural investment in Ghana. *J. Political Econ.* 116 (6), 981–1022.
- Hancock, J., 1992. Extension Education: Conducting Effective Agricultural Demonstrations. Cooperative Extension Service. College of Agriculture, University of Kentucky.
- Hanna, R., Mullainathan, S., Schwartzstein, J., 2014. Learning by noticing: theory and evidence through a field experiment. *Q. J. Econ.* 129 (3), 1311–1353.
- Hornbeck, R., Keskin, P., 2015. Does agriculture generate local economic spillovers? Short run and long run evidence from the Ogallala aquifer. *Am. Econ. J. Econ. Policy* 7 (2), 192–213.
- Islam, M., 2014. Practice Does Not Make Perfect: Understanding Fertilizer Mismanagement in Bangladesh through Leaf Color Charts. Working Paper. Harvard University.
- Jack, K., 2011. Market Inefficiencies and the Adoption of Agricultural Technologies in Developing Countries. White paper prepared for the Agricultural Technology Adoption Initiative (ATAI), JPAL (MIT)/CEGA, Berkeley.
- Jack, K., 2013. Private information and the allocation of land use subsidies in Malawi. *Am. Econ. J. Appl. Econ.* 5 (3), 113–135.
- Jack, W., de Laat, J., Kremer, M., Suri, T., 2015. Joint Liability, Asset Collateralization, and Credit Access: Evidence from Rainwater Harvesting Tanks in Kenya (Working Paper).
- Jacoby, H., 1993. Shadow wages and peasant family labor supply: an econometric application to the Peruvian Sierra. *Rev. Econ. Stud.* 60, 903–921.
- Jayachandran, S., 2006. Selling labor low: wage responses to productivity shocks in developing countries. *J. Political Econ.* 114 (3), 538–575.
- Kala, N., 2014. Ambiguity Aversion and Learning in a Changing World: The Potential Effects of Climate Change from Indian Agriculture. Working Paper. Yale University.
- Karlan, D., Osei, R., Osei-Akoto, I., Udry, C., 2014. Agricultural decisions after relaxing credit and risk constraints. *Q. J. Econ.* 129 (2), 597–652.
- Lichtenberg, E., Zilberman, D., 1986. The econometrics of damage control: why specification matters. *Am. J. Agric. Econ.* 68 (2), 261–273.
- Mahajan, A., Seira, E., Gine, X., 2014. A Multiple Interventions Approach to Increasing Technology Adoption with a View towards Scaling-up: Evidence from Mexico (BASIS Proposal).

- Matsumoto, T., Yamano, T., Sserunkuuma, D., 2013. Technology Adoption and Dissemination in Agriculture: Evidence from Sequential Intervention in Maize Production in Uganda. GRIPS Discussion Papers 13–14. National Graduate Institute for Policy Studies.
- McIntosh, C., 2014. Building Markets for Small Scale Farmers. [http://cega.berkeley.edu/events/E2A\\_2014/](http://cega.berkeley.edu/events/E2A_2014/).
- McIntosh, C., Sarris, A., Papadopoulos, F., 2013. Productivity, credit, risk, and the demand for weather index insurance in smallholder agriculture in Ethiopia. *Agric. Econ.* 44, 399–417.
- Mobarak, M., Rosenzweig, M., 2014. Risk, Insurance and Wages in General Equilibrium. Working Paper. Yale University.
- Mobarak, M., Rosenzweig, M., 2013. Selling Formal Insurance to the Informally Insured. Working paper. Yale University.
- Renkow, M., Hallstrom, D., Karanja, D., 2004. Rural infrastructure, transactions costs and market participation in Kenya. *J. Dev. Econ.* 73 (1), 349–367.
- Sadoulet, E., de Janvry, A., 1995. Quantitative Development Policy Analysis. The Johns Hopkins University Press, Baltimore.
- Schickele, A., 2016. “Make it rain.” *Policy Bulletin*, February, J-PAL, CEGA, ATAI.
- Schilbach, F., 2015. Essays in Development and Behavioral Economics. (Doctoral Dissertation) Harvard University, Graduate School of Arts & Sciences.
- Schultz, T.W., 1964. Transforming Traditional Agriculture. Yale University Press, New Haven.
- Suri, T., 2011. Selection and comparative advantage in technology adoption. *Econometrica* 79 (1), 159–209.
- Tjernström, E., 2014. Signals, Similarity and Seeds: Social Learning in the Presence of Imperfect Information and Heterogeneity. Working Paper. University of California at Davis.
- Visaria, S., Mitra, S., Mookherjee, D., Torero, M., 2015. Asymmetric Information and Middleman Margins: An Experiment with Indian Potato Farmers. HKUST IEMS Working Paper No. 2015–29.
- World Bank, 2007. Agriculture for Development–World Development Report. The World Bank.
- World Bank, 2016. World Bank Development Indicators. Online Database. The World Bank.
- Zezza, A., Federighi, G., Adamou, K., Hiernaux, P., 2014. Milking the Data: Measuring Income from Milk Production in Extensive Livestock Systems: Experimental Evidence from Niger. Research working paper WPS 7114. World Bank Group, Washington, DC.

## CHAPTER 6

# The Personnel Economics of the Developing State

F. Finan<sup>\*</sup>, B.A. Olken<sup>§,1</sup>, R. Pande<sup>¶,a</sup>

<sup>\*</sup>University of California, Berkeley, Berkeley, CA, United States

<sup>§</sup>Massachusetts Institute of Technology, Cambridge, MA, United States

<sup>¶</sup>Harvard University, Cambridge, MA, United States

<sup>1</sup>Corresponding author: E-mail: bolken@mit.edu

## Contents

1. Introduction	468
2. Stylized Facts on the Architecture of the State and the Role of Individuals	470
2.1 Key features of the state	470
2.2 Evidence from household surveys	472
3. The Selection and Recruitment of Public Officials	482
3.1 Financial incentives	485
3.1.1 Effects of financial incentives on the applicant pool	485
3.1.2 Effects of financial incentives on recruitment	487
3.2 How should governments screen?	488
4. Using Incentives to Improve Performance	491
4.1 Financial incentives	492
4.1.1 Incentives for agents of government authority	492
4.1.2 Incentives for frontline service providers	493
4.2 Nonfinancial incentives	497
4.2.1 Transfers and postings	497
4.2.2 Intrinsic motivation	498
4.3 Summary	499
5. Monitoring Mechanisms and Public Service Delivery	500
5.1 Overview	500
5.2 Information flows and monitoring	500
5.3 Government monitoring	500
5.3.1 Does more information on performance improve outcomes?	500
5.3.2 Who collects information and does that matter?	503

<sup>a</sup> The authors thank Nils Enevoldsen and Joyce Hahn for exceptional research assistance, and Abhijit Banerjee, Alan Gerber, and Adnan Khan for comments. The authors acknowledge financial support from the William and Flora Hewlett Foundation and the UK Department for International Development. The views expressed here are those of the authors alone.

5.4 Information flows and monitoring by citizens	504
5.4.1 Does information on program performance matter?	504
5.5 Summary	505
6. Towards Smart(er) Governance: the Promise of e-Governance and Other Avenues	505
7. Concluding Thoughts	507
Appendix	509
References	511

## Abstract

Governments play a central role in facilitating economic development. Yet while economists have long emphasized the importance of government quality, historically they have paid less attention to the internal workings of the state and the individuals who provide the public services. This chapter reviews a nascent but growing body of field experiments that explores the personnel economics of the state. To place the experimental findings in context, we begin by documenting some stylized facts about how public sector employment differs from that in the private sector. In particular, we show that in most countries throughout the world, public sector employees enjoy a significant wage premium over their private sector counterparts. Moreover, this wage gap is largest among low-income countries, which tends to be precisely where governance issues are most severe. These differences in pay, together with significant information asymmetries within government organizations in low-income countries, provide a *prima facie* rationale for the emphasis of the recent field experiments on three aspects of the state-employee relationship: selection, incentive structures, and monitoring. We review the findings on all three dimensions and then conclude this survey with directions for future research.

## Keywords

Financial incentives; Public sector; Public sector wage gap; Public service delivery or public goods; State capacity

## JEL Codes

H40; M50; O10; O43

## 1. INTRODUCTION

Countries vary in their quality of governance, and by almost any measure, governance is significantly worse in low-income countries. For instance, the World Bank's Worldwide Governance Indicators project rank low-income countries substantially lower than the high-income countries that are members of the Organization for Economic Co-operation and Development (OECD) on government effectiveness (average percentile rank of 17.3 compared to 87.9 in 2014). For frontline service providers, such as teachers and nurses, a well-known study of six developing countries found that 19 percent of public primary school teachers and 35 percent of public health care workers were absent at the time of random unannounced visits to schools and clinics. The absenteeism was worse in poorer countries and in poorer states within India, with the worst absenteeism rates approaching 40 percent ([Chaudhury et al., 2006](#)). Given the natural role governments play in facilitating development, whether by providing public goods, addressing

externalities, or providing the foundation for private property and private enterprise, improvements in government performance are likely to lead to significant economic development gains.

In this chapter, we examine a particular determinant of government performance: the individuals who perform government functions. We focus on two groups of public employees: appointed civil servants, which we broadly construe to include administrators with effectively permanent government appointments and frontline service providers (e.g., teachers, nurses, firefighters, and trash collectors), who may have either permanent appointments or temporary contracts<sup>1</sup>.

In many developing countries, policy actors and researchers attribute poor governance to public employees being lazy, corrupt, or both. Yet, it may be that the poor institutional structure within which public employees work in these countries is the dominant cause. The two may also interact: poor institutional structures may cause the lazy and the corrupt to select into public service. Since both institutional structures and personnel selection are endogenously determined, establishing causality is hard. By providing a clean empirical method for identification, field experiments can help cut the Gordian knot and identify problems in personnel selection and management, elucidate the causes of these problems, and suggest potential solutions. These contributions of field experiments are the focus of this chapter.

While the focus of this chapter is the “personnel economics” of the government sector—in particular, selection and recruitment of personnel, incentives, and monitoring—there are many aspects of the government as an organization that are also clearly important but not covered by this review.<sup>2</sup> In particular, a recent literature has emphasized the role of management skills and techniques, which are conceptually distinct from the people who implement them (e.g., [Bloom and Van Reenen, 2007](#); [Bloom et al., 2013](#); [Rasul and Rogger, 2015](#)). While these are key issues in determining the overall “Total Factor Productivity” (TFP) of the government sector, they are not the issues we explicitly consider here.

The chapter is organized as follows. In [Section 2](#), we situate the problem by documenting several stylized facts of how the public sector as an employer differs from the private sector. We use survey microdata from 32 countries to establish some facts about the relative compensation and fringe benefits of public and private sector workers. The per capita Gross Domestic Product (GDP) of our study countries ranges from \$264 to \$45,710 (constant 2005 USD). The World Bank classifies 19 of our sample countries as low income or lower middle income. We use these surveys to document the stark difference between the public–private gaps in worker pay and worker tenure in developing

<sup>1</sup> The occupational categories of contract workers are typically varied, but important categories in our empirical sample include teachers, nurses, office cleaners, and helpers.

<sup>2</sup> For a review of personnel economics with a focus on the private sector, see, for example, [Lazear and Oyer \(2012\)](#).

versus developed countries: the public sector enjoys a large wage premium in poor countries but only a small—or, in some cases, zero—premium in developed countries. Public sector jobs are also more likely to provide fringe benefits like health insurance and pensions; again, the difference is much starker in poor as opposed to rich countries.

These stylized facts point to a large pay premium for public sector employees in low-income countries. The premium could reflect the more complex nature of public sector jobs in low-income countries, elite capture of the public sector, efficiency wages designed to prevent corruption a la [Becker and Stigler \(1974\)](#), differences in job preferences across societies, or some combination of these factors. Whatever the underlying cause may be, this pay premium—in combination with weak information flows within and across government agencies—has important implications for how individuals select into the public sector and their subsequent performance and incentive structure. Against this background, we argue that field experiments can play an important role in helping us understand the links between governance outcomes and one dimension of state capacity: the traits and behavior of public employees.

[Sections 3–5](#) then describe how field experiments have informed our understanding of the current problems in three arenas (selection and recruitment, incentives, and monitoring) and helped identify potential civil service reforms. In [Section 6](#), we discuss whether the changing nature of technology-driven aids for service delivery may help poor countries create smart governance structures, perhaps by constraining or eliminating human interaction. We conclude with directions for future research.

## 2. STYLIZED FACTS ON THE ARCHITECTURE OF THE STATE AND THE ROLE OF INDIVIDUALS

### 2.1 Key features of the state

The state, the world over, consists of a set of interlinked institutions staffed by officials. Typically, a mix of constitutional acts and legislative and executive orders defines the fiscal and regulatory powers of these institutions. In the Weberian model, adopted in virtually all countries (though to varying degrees; see [Evans and Rauch, 1999](#)), apolitical civil servants, in turn, are responsible for implementing the state mandate.

Several key features of the state distinguish its personnel practices from those of the private sector, particularly in developing countries, and here, we identify five features that are relevant for our analysis.

First, the state has a long horizon. Absent significant civil conflict, most states anticipate collecting and spending revenues indefinitely. This allows states to make long-lived promises to its employees, such as pensions, which may be difficult for the private sector to make. The long-lived reputation of the state as an employer also means that it may be reluctant to renege on such promises; in fact, many promises of the state, again such as pensions, survive radical regime change, at least in nominal terms.

Second, the set of contracts a state can offer its employees is limited. Whereas shareholders can create strong incentives for CEOs to maximize returns, the mechanisms that the ultimate principals—citizens—have at their disposal, namely elections, are coarser and more limited. Thus, politicians may seek to use jobs and the wages associated with them, to reward their political supporters, cronies, and friends. Politicians could also be tempted to use promotions or incentives to exert undue influence on civil servants. These issues are more prevalent in the public sector than in the private sector because of the lack of discipline from the profit motive: the politician only indirectly bears the cost of inefficiency to the extent that voters are less likely to re-elect him because of it, whereas the owner of a firm directly feels the financial losses associated with these types of inefficiencies (Shleifer and Vishny, 1994; Boycko et al., 1996). To counteract the tendency for each new politician to replace large numbers of government employees with his political supporters, over time governments have enacted rigid civil service rules that restrict the discretion politicians have over hiring and firing (Evans, 1995 AER).<sup>3</sup> In fact, these civil service systems are typically much more rigid than their private sector counterparts, with strict formulas defining the hiring criteria, promotion patterns, and wage levels. The need to isolate the employment decision from political influence—which underlies a substantial amount of public sector personnel policy—suggests that the personnel economics of the state are likely to substantially differ from those of the private sector.

Another restriction on contracts is that public sector compensation usually does not include pay for performance. Performance pay for bureaucrats can create severe multitasking problems, where bureaucrats focus on the incentivized dimension of their job at the expense of the nonincentivized dimension (Holmstrom and Milgrom, 1987). While multitasking is an issue in many contexts, it can be particularly severe in public sector contexts where agents wield substantial authority (e.g., police and judges), and it is hard to find an objective measure of the “truth” on which to incentivize them. In practice, while financial incentives for government workers were historically quite common, they had a tendency to lead to overzealous and unpopular bureaucrats who were perceived to abuse their positions in order to overextract from the population (e.g., Parrillo, 2013), which led the populace to demand less strongly incentivized civil servants.

Third, the nature of goods exchanged between the state and citizens is substantially different than with the private sector. Very often, services provided by the state—like schooling and health—are heavily subsidized, thus limiting the competition the state faces

<sup>3</sup> The United States at the turn of the 19th century provides a nice illustration. Prior to the Pendleton Act, federal employment was an important source of patronage. Under this system, these employees did not have tenure and turnover rates were high during changes in administration. After the passage of the Pendleton Act, which restricted the number of patronage positions, civil service reform started to take hold and federal employees began to get hired based on merit and public service exams. With these reforms, the job took on a different form. Federal employees were granted tenure and dismissals became more difficult and costly. Compensation became more formulaic which resulted in more wage compression.

from other providers. Because there is little competition from other providers, it becomes harder to base worker incentives on simple metrics like volume of services, and the lack of competitive pressure makes direct monitoring of service providers more important for the public sector (relative to the private sector where competition will naturally weed out less productive firms).

Fourth, government careers differ from nongovernment careers in the mission of the organization: government organizations often aspire to public service and private sector ones to profit. This, arguably, implies that different types of individuals are potentially drawn to the public and private sector careers, and personnel practices should account for this. The state may also seek different types of individuals who are able to balance the multifaceted objective functions inherent to a public service organization. [For a theoretical discussion of how matching mission organization to agent preferences can improve efficiency, see [Besley and Ghatak \(2005\)](#).]

Finally, a fifth key feature of the state is that it self-regulates to a much larger extent than the private sector. Monitoring structures are often embedded within the bureaucracy, and there are relatively few instances of third-party private auditing of government services. Moreover, in several cases, workers often switch between service delivery and monitoring roles, which potentially leads to conflicts of interest.

## 2.2 Evidence from household surveys

To quantify some of these differences between the public and private sector, we obtained household survey microdata from 32 countries around the world. [Appendix Table 1](#) lists these countries and the data sources.

We classify each working adult as a public or private sector employee using information provided in the survey specifying the type of employer. For these adults, we also know their wages and usually their occupations and the number of years at their primary jobs. In terms of other forms of employer-provided compensation, we have information on pensions for 13 countries and on health insurance for 14 countries. (We exclude state-provided benefits which may be available to all citizens, irrespective of work status.) In terms of worker demographics, in addition to occupation, we know gender, age, and education level.

We begin by estimating the public sector wage premium, separately for each country  $c$ , as follows:

$$y_{irc} = \alpha_r + \beta_c Public_{irc} + \gamma_c X_{irc} + \epsilon_{irc}$$

where  $i$  indexes an individual,  $c$  indexes a country,  $r$  indexes a region within a country,  $Public_{irc}$  is a dummy for public sector, and  $X_{irc}$  are vector of controls (age, gender, secondary education, tertiary education, and occupation dummies). We begin with log wages as the outcome variable but also consider other aspects of compensation, such as pensions, health benefits, and job tenure. The key coefficient of interest is  $\beta_c$ ,

which captures, separately for each country  $c$ , the differential return for being in the public sector.

[Table 1](#) reports the results where the outcome variable is log wages. Each cell reports the coefficient  $\beta_c$  from a separate regression. Column (1) reports the basic model with a single constant term for each country (i.e., no regional fixed effects) and no control variables. Column (2) adds regional (usually, province level) dummies for each country to capture geographic differences; for example, government workers may be disproportionately located in the capital city, which may have different wage levels.<sup>4</sup> Column (3) adds as covariates age, gender, and education. Column (4) adds occupation fixed effects (e.g., in Uganda, commonly observed occupations included mechanics, nurses, and midwives, managing supervisors, transport laborers, restaurant service workers, and machinery mechanics and fitters). For many occupations there is substantial overlap between public and private sector. Teachers, for example, work in both sectors, as do many types of service workers. We show the results separately for each country.

In order to facilitate interpretation of the results, [Fig. 1](#) plots the estimated coefficient  $\beta$ , the public sector premium, against each country's 2010 per capita GDP, measured in terms of Purchasing Power Parity (PPP). The left panel plots the coefficients from column (1), the unadjusted public–private wage difference. The right panel plots the coefficients from column (4), that is, after including region fixed effects, occupation fixed effects, and demographic controls. Since the public sector was substantially different in communist economies, we plot two regression lines: the red line shows a regression line for current and former communist countries, and the blue line shows the regression line for all other countries.

The left graph shows that, in almost all countries unadjusted average pay in the public sector exceeds that in the private sector. The difference is much starker in poor countries: in the poorest countries in our sample, such as Malawi, Niger, Tanzania, and Kenya, public sector workers earn more than double the average wage in the private sector. This declines to the point where for the rich countries in our sample—Korea, the United Kingdom, and the United States—the pay difference, while present, is on the order of 4–20 percent.

The right graph in [Fig. 1](#) plots the difference after we include all the controls discussed above. For nonformer communist countries, we still see a negative relationship between income and public sector wages, but it is muted substantially; the poorest four countries in the sample have a positive public sector premium of around 0.1, whereas the richest three countries in the sample have a negative public sector premium of around –0.03. Comparing the results in each column, the key difference from the unadjusted results

<sup>4</sup> Whenever possible, we used enumeration area codes or primary sampling unit codes for geographic fixed effects. For the countries that did not have those, we grouped multiple geographic identifiers together. For example, for the case of Argentina, we grouped region and agglomeration identifiers. For South Africa, we grouped province and district council. For the United States, we grouped region, FIPS state code, and metropolitan CBSA FIPS code.

**Table 1** Log pay on public sector: four models of increasing specification

	(1)	(2)	(3)	(4)
	Basic model	Add region fixed effects	Add individual demographic adjustments	Add occupation fixed effects
Albania	0.077* (0.044)	0.124* (0.068)	0.017 (0.080)	-0.001 (0.095)
Argentina	0.3468*** (0.012)	0.373*** (0.011)	0.241*** (0.012)	0.041*** (0.013)
Armenia	-0.098** (0.050)	-0.157** (0.062)	-0.061 (0.061)	-0.073 (0.074)
Bolivia	0.202*** (0.078)	0.176* (0.095)	-0.042 (0.101)	-0.077 (0.126)
Bosnia and Herzegovina	0.037 (0.025)	0.043* (0.025)	-0.051** (0.025)	-0.119*** (0.027)
Bulgaria	-0.076*** (0.017)	-0.061*** (0.019)	-0.043** (0.019)	-0.059*** (0.021)
Colombia	0.633*** (0.109)	0.457*** (0.137)	0.302** (0.134)	0.127 (0.148)
Egypt	-0.150*** (0.017)	-0.102* (0.054)	-0.274*** (0.061)	-0.227* (0.124)
Georgia	-0.000 (0.061)	0.083 (0.062)	0.137** (0.060)	0.243*** (0.078)
Ghana	0.937*** (0.089)	0.956*** (0.129)	0.588*** (0.136)	0.764*** (0.187)
India	1.113*** (0.009)	1.101*** (0.010)	0.712*** (0.010)	0.641*** (0.012)
Indonesia	0.847*** (0.032)	0.953*** (0.033)	0.462*** (0.036)	0.546*** (0.039)
Iraq	0.162*** (0.009)	0.127*** (0.011)	0.108*** (0.012)	-0.039** (0.018)
Kenya	0.886*** (0.107)	0.770*** (0.125)	0.473*** (0.116)	0.419*** (0.140)
Korea, Rep.	0.040 (0.101)	0.093 (0.112)	0.110 (0.095)	-0.224 (0.208)
Laos	-0.265*** (0.053)	-0.182*** (0.063)	-0.375*** (0.079)	-0.483*** (0.123)
Malawi	0.800*** (0.039)	0.700*** (0.046)	0.234*** (0.049)	0.126** (0.054)
Mexico	0.477*** (0.005)	0.496*** (0.005)	0.433*** (0.006)	0.220*** (0.007)
Nicaragua	0.502*** (0.028)	0.455*** (0.028)	0.262*** (0.081)	0.153 (0.130)
Niger	0.986*** (0.088)	0.676*** (0.120)	0.249** (0.118)	-0.056 (0.149)

**Table 1** Log pay on public sector: four models of increasing specification—cont'd

	(1)	(2)	(3)	(4)
	Basic model	Add region fixed effects	Add individual demographic adjustments	Add occupation fixed effects
Nigeria	0.719*** (0.059)	0.669*** (0.083)	0.355*** (0.086)	0.359*** (0.106)
Pakistan	0.540*** (0.013)	0.541*** (0.014)	0.258*** (0.013)	0.322*** (0.014)
Panama	0.637*** (0.020)	0.574*** (0.022)	0.465*** (0.022)	0.388*** (0.023)
Peru	0.296*** (0.030)	0.188*** (0.034)	0.065** (0.032)	0.068* (0.038)
Serbia	0.419*** (0.019)	0.333*** (0.019)	0.249*** (0.019)	0.096*** (0.022)
Sri Lanka	0.437*** (0.047)	0.341*** (0.061)	0.347*** (0.069)	0.252*** (0.102)
Tajikistan	0.061 (0.048)	-0.201*** (0.056)	-0.193*** (0.057)	-0.317*** (0.091)
Tanzania	0.962*** (0.057)	0.962*** (0.066)	0.483*** (0.078)	0.285*** (0.097)
Timor Leste	0.325*** (0.075)	0.159* (0.096)	0.008 (0.101)	-0.086 (0.101)
Uganda	0.546*** (0.088)	0.611*** (0.108)	0.051 (0.111)	0.035 (0.139)
United Kingdom	0.086*** (0.015)	0.094*** (0.015)	0.036** (0.014)	0.025 (0.017)
United States	0.216*** (0.019)	0.216*** (0.019)	0.113*** (0.019)	0.108*** (0.020)
Vietnam	0.120** (0.052)	0.040 (0.060)	-0.082 (0.058)	-0.066 (0.060)

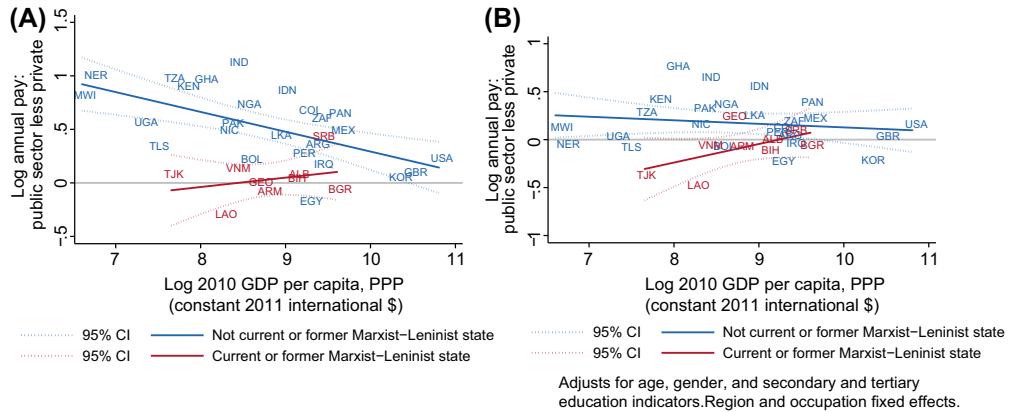
Standard errors in parentheses.

Each table cell contains a within-country estimate of the difference in log pay between the public sector and the private sector. Ambiguous sectors such as NGO are omitted. Estimates are obtained with Ordinary Least Squares (OLS) regression. Column 1 is the basic specification. To column 1, column 2 adds region fixed effects, such as district or municipality. To column 2, column 3 adds covariates for age, gender, and indicators for completion of secondary education and for completion of any tertiary education.

To column 3, column 4, the fully specified model, adds occupation fixed effects, as determined by occupation codes. Each country's data come from a single household-level survey, as listed in the [Appendix](#).

\*,  $p < 0.10$ ; \*\*,  $p < 0.05$ ; \*\*\*,  $p < 0.01$ .

comes when we add controls for education in column (3), an issue we will return to in more detail in the following section. For ex-communist countries, there is an upward slope, so that the government workers in poor former or current communist countries appear to have a substantial negative wage premium relative to their private sector counterparts.



**Figure 1** Public sector pay premium by GDP per capita. (A) Public sector pay premium, basic model. (B) Public sector pay premium, fully specified model. *Cl*, Confidence interval.

Table 2 and Figs. 1–3 show the results for other outcome variables, notably whether the worker receives health insurance, whether he receives a pension, and his tenure on the job. The public sector looks remarkably different from the private sector in these other benefits in both rich and poor countries. In particular, on average, public sector workers are about 20 percentage points more likely to receive health insurance than private sector workers, even conditional on other job characteristics, and this does not systematically differ between rich and poor countries. In some countries, these differences are even greater; in India, for example, the public sector is 48 percentage points more likely to receive health insurance and 55 percentage points more likely to receive pensions, even conditional on job characteristics. These fringe benefits, which may reflect the government's ability to honor commitments across states of the world (e.g., if a worker falls ill) or over time (i.e., when a worker grows old) given its much longer time horizons, are a notable difference between the public and private sector.

A different type of benefit, which we focus on in Fig. 4, is job tenure. We observe a significant public sector premium: in the unadjusted model public sector workers report having had that job for 5 years longer than their private sector counterparts. In the adjusted model, we continue to see a significant positive premium of roughly 3 years. There is no trend in this premium by country income.

Given the differences between the adjusted and unadjusted wage regressions, we next examine the differences in observable demographic characteristics. In Fig. 5, we examine worker gender. (The figure plots country-wise coefficients on the public sector dummy where the estimation equation is of the form in column (1) of Table 1.) On average, countries have a relatively higher fraction of women in the public sector relative to the private sector, and feminization of the public sector relative to private sector is

**Table 2** Job benefits on public sector

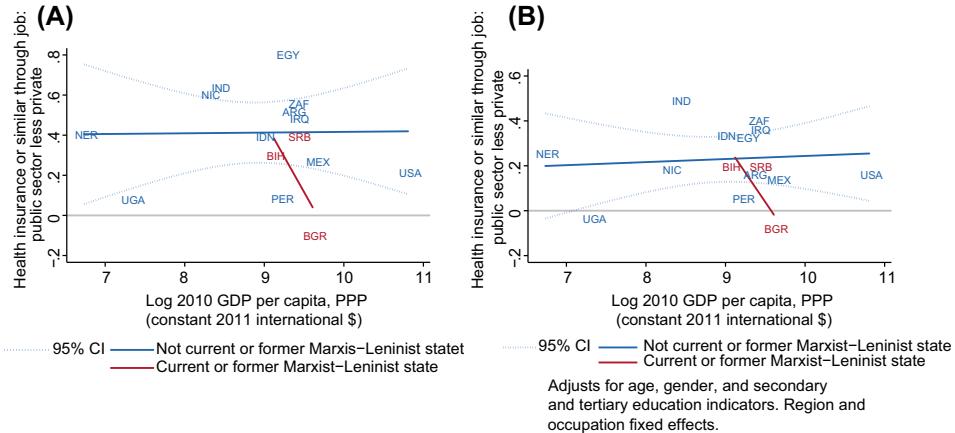
	(1)	(2)	(3)
	Log pay	Health benefits	Pension
Argentina	0.041*** (0.013)	0.155*** (0.008)	0.146*** (0.009)
Bosnia and Herzegovina	-0.119*** (0.027)	0.194*** (0.022)	0.208*** (0.022)
Bulgaria	-0.059*** (0.021)	-0.084*** (0.011)	-0.084*** (0.011)
Egypt	-0.227* (0.124)	0.334*** (0.064)	—
India	0.641*** (0.012)	0.483*** (0.007)	0.545*** (0.007)
Indonesia	0.546*** (0.039)	0.330*** (0.017)	0.479*** (0.016)
Iraq	-0.039** (0.018)	0.346*** (0.011)	0.603*** (0.011)
Mexico	0.220*** (0.007)	0.133*** (0.004)	0.133*** (0.004)
Nicaragua	0.153 (0.130)	0.176** (0.085)	-0.214** (0.084)
Niger	-0.056 (0.149)	0.249*** (0.061)	0.286*** (0.076)
Nigeria	0.359*** (0.106)	—	0.555*** (0.055)
Peru	0.068*** (0.038)	0.048 (0.033)	—
Serbia	0.096*** (0.022)	0.190*** (0.012)	0.402*** (0.026)
South Africa	0.185*** (0.015)	0.374*** (0.006)	0.259*** (0.005)
Uganda	0.035 (0.139)	-0.040 (0.063)	0.047 (0.071)
United States	0.108*** (0.020)	0.157*** (0.012)	0.204*** (0.012)

Standard errors in parentheses.

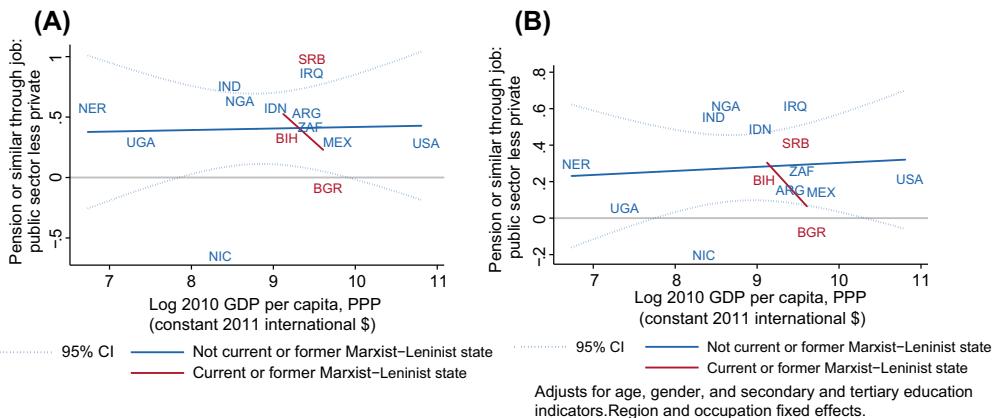
Each table cell contains a within-country estimate of the difference in benefits between the public sector and the private sector. Ambiguous sectors such as NGO are omitted. Estimates are obtained with OLS regression, using the fully specified model described in Table 1. Estimates have region and occupation fixed effects, as well covariates for age, gender, and indicators for completion of secondary education and for completion of any tertiary education. The dependent variable in column 1 is log pay and is identical to column 4 of Table 1. The dependent variable of column 2 is an indicator for employer-provided health benefits, such as insurance. The dependent variable for column 3 is an indicator for employer-provided pension. Each country's data come from a single household-level survey, as listed in the Appendix.

Albania and Mexico have single variables for health and pension; these estimates are duplicated across health and pension columns. Nicaragua's pension variable indicates if employer offers any pension benefits, while the health variable asks about health benefits in addition to but not separate from pension benefits.

\*,  $p < 0.10$ ; \*\*,  $p < 0.05$ ; \*\*\*,  $p < 0.01$ .



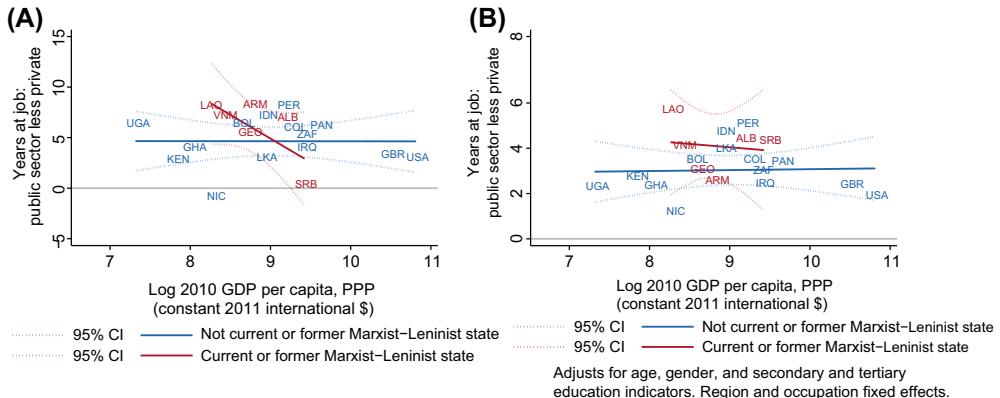
**Figure 2** Public sector health benefit premium by GDP per capita. (A) Public sector health benefit premium, basic model. (B) Public sector health benefit premium, fully specified model. *CI*, confidence interval.



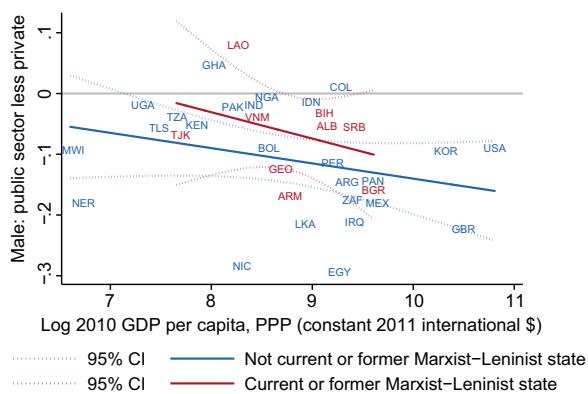
**Figure 3** Public sector pension premium by GDP per capita. (A) Public sector pension premium, basic model. (B) Public sector pension premium, fully specified model. *CI*, confidence interval.

increasing with income. Fig. 6 considers age, and we find that public sector workers tend to be, on average, 5 years older. There is no significant gradient with country income.

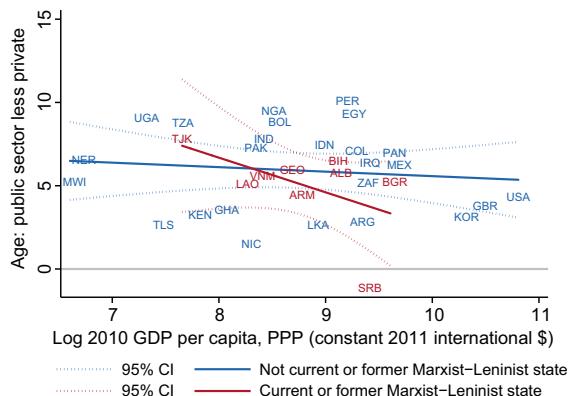
In Fig. 7, we turn to education. The variable of interest is an indicator of whether the worker has completed at least secondary education. We observe both a public sector education premium, wherein more educated workers are attracted to the public sector at a higher rate and a significant negative gradient in this education premium by county



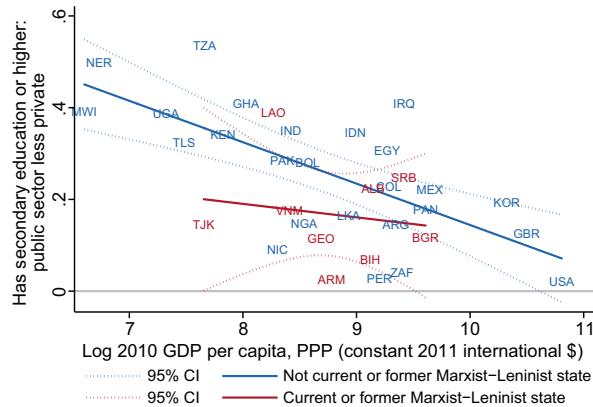
**Figure 4** Public sector tenure premium by GDP per capita. (A) Public sector tenure premium, basic model. (B) Public sector tenure premium, fully specified model. *CI*, confidence interval.



**Figure 5** Public sector gender difference by GDP per capita. *CI*, confidence interval.



**Figure 6** Public sector age difference by GDP per capita. *CI*, confidence interval.



**Figure 7** Public sector education difference by GDP per capita, unadjusted. *Cl*, confidence interval.

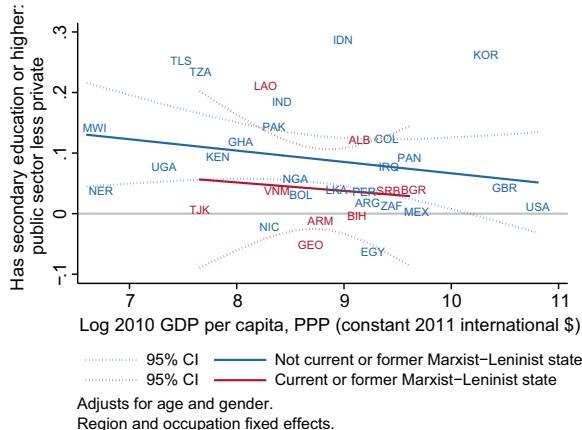
income. This strong gradient in the education premium, together with its much weaker fully specified counterpart in Fig. 8, underlies the significant difference in the income gradient in public sector wage premium across panels (A) and (B) of Fig. 1.

Finally, in Fig. 9, we report estimates based on a basic Mincer wage regression where we report the differential public sector wage premium for years of education. Differential Mincerian educational premia are low and show no income gradient.

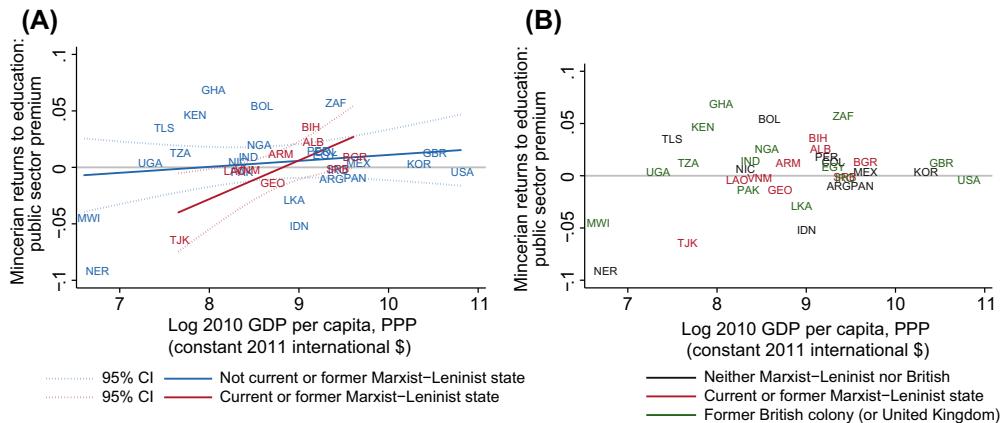
The survey data paint an intriguing picture: In poorer countries, the public sector attracts relatively more educated individuals, who in turn are paid an equivalent wage and, in almost all places, higher wage premium. Yet, qualitative evidence suggests that the quality of government services is lower in developing countries than in developed ones. While there are a myriad of explanations for poor government performance, the evidence presented here suggests that—assuming similar job descriptions across countries—one can likely rule out the explanation that government workers are underpaid relative to their private sector counterparts.<sup>5</sup> Rather, it is likely that, in settings with weak information flows within the government and between the government and citizens, high wages per se are insufficient to motivate performance (though we suggest further direct evidence on this relationship in the following section).

The key point established here is that the public sector is substantially different, in terms of the level of wages, composition of the labor force, fringe benefits, and tenure. But understanding the impact of these personnel policies—and ways they can be altered

<sup>5</sup> We, of course, do not know the counterfactual. It may be that government performance in developing countries would be even worse if pay was lower, if high pay is acting as an efficiency wage to prevent corruption. Alternatively, it may be that the job description is more taxing in low-income countries, causing the effective wage to be lower. While we cannot directly rule this out, the evidence on, for instance, teacher and health worker absenteeism across rich and poor countries would go against such an explanation.



**Figure 8** Public sector education difference by GDP per capita, adjusted. *Cl*, confidence interval.



**Figure 9** Mincerian returns to education, public sector premium. (A) Mincerian returns to education, public sector premium. (B) Mincerian returns to education, public sector premium. *Cl*, confidence interval.

to improve government performance—is challenging. Field experiments offer an attractive way of providing evidence on these outcomes, by examining what happens when these practices are altered or changed on various dimensions. To explore these issues in detail and shed light on how to think about various aspects of personnel policy in the government sector, in the remaining sections, we explore the evidence on three dimensions through which performance of government employees is determined and potentially can be improved: selection and recruitment (Section 3), incentives (Section 4), and monitoring (Section 5). In each case, we describe existing evidence (focused, in particular, on field experiments) and point out important open questions for future

research. The first section (selection and recruitment) considers how to improve the selection of public employees with the set of attributes that best contribute to public sector performance; the next two sections (incentives and monitoring) consider how to improve the performance of a given set of individuals, holding their basic characteristics fixed.

### **3. THE SELECTION AND RECRUITMENT OF PUBLIC OFFICIALS**

The first questions we consider are: who are government employees, how are they recruited, and are there ways of improving the recruitment process? After all, individuals choose their career paths given the options available to them. In addition to the differences in compensation and job tenure documented previously, a key difference between government and nongovernment careers that we identified earlier is mission differences: government organizations often aspire to public service and private sector ones to profit. On the other hand, the rampant corruption in the public sector in developing countries may attract those who are interested in pursuing corrupt activities. More generally, [Table 3](#) highlights several descriptive differences between the public and private sector across countries.

Do mission-driven organizations, such as public bureaucracies or private nonprofit organizations, attract employees with high levels of prosocial motivation? The idea is that some individuals care about benefiting others and thus feel drawn to organizations that provide them with the opportunity to do so. In general, the literature in public administration and economics supports this idea. For example, [Cowley and Smith \(2014\)](#) use data from the World Values Survey to measure the intrinsic motivation of public and private sector employees in 52 countries. They find that public sector workers are on average much more intrinsically motivated than private sector workers, even after adjusting for differences in basic socioeconomic characteristics. Similarly, [Banuri and Keefer \(2013\)](#) sampled about 1700 individuals from the government and private sectors in Indonesia and had them play a dictator game. They also find that subjects in the governmental sector are more prosocial.<sup>6</sup>

Besides the prosocially motivated, public bureaucracies can also attract individuals with less desirable personality traits. Organizations that offer low-powered incentives or are unable to hold their employees accountable can attract individuals with limited aspirations and a poor work ethic. Widespread corruption may attract dishonest or venal individuals. Recent evidence also supports this view. In the study by [Cowley and Smith \(2014\)](#), for example, although public sector workers tend to be more intrinsically

<sup>6</sup> Other examples include: [Dohmen and Falk \(2010\)](#) find that German teachers are more trusting and less negatively reciprocal than employed nonteachers. [Lagarde and Blaauw \(2014\)](#) find in an adapted dictator game that giving to patients predicts student nurses' subsequent decisions to take rural hardship posts in South Africa. See [Perry and Hondeghem \(2008\)](#) for additional studies examining the role of prosocial motivation in selection in public service.

**Table 3** Demographic differences of public sector

	(1)	(2)	(3)	(4)
	Male	Age	Secondary or tertiary education	Tertiary education
Albania	-0.039*	5.748*** (0.024)	0.224*** (0.018)	0.320*** (0.021)
Argentina	-0.147*** (0.008)	2.832*** (0.203)	0.145*** (0.006)	0.258*** (0.008)
Armenia	-0.169*** (0.037)	4.426*** (0.973)	0.024* (0.014)	0.249*** (0.034)
Bolivia	-0.090** (0.044)	8.816*** (0.978)	0.279*** (0.025)	0.448*** (0.032)
Bosnia and Herzegovina	-0.028 (0.023)	6.460*** (0.502)	0.068*** (0.019)	0.175*** (0.017)
Bulgaria	-0.159*** (0.017)	5.473*** (0.374)	0.117*** (0.011)	0.274*** (0.016)
Colombia	0.010 (0.090)	6.900*** (1.807)	0.228*** (0.046)	0.359*** (0.084)
Egypt	-0.295*** (0.009)	9.283*** (0.218)	0.306*** (0.009)	0.314*** (0.010)
Georgia	-0.125*** (0.038)	5.918*** (0.932)	0.113*** (0.030)	0.125*** (0.034)
Ghana	0.047 (0.049)	3.395*** (1.021)	0.409*** (0.040)	0.450*** (0.048)
India	-0.019*** (0.005)	7.785*** (0.122)	0.350*** (0.005)	0.301*** (0.005)
Indonesia	-0.015 (0.015)	7.442*** (0.315)	0.346*** (0.010)	0.413*** (0.014)
Iraq	-0.211*** (0.004)	6.647*** (0.162)	0.408*** (0.006)	0.342*** (0.006)
Kenya	-0.048 (0.060)	3.234*** (1.064)	0.340*** (0.028)	0.366*** (0.059)
Korea, Rep.	-0.087 (0.068)	3.144* (1.642)	0.192*** (0.060)	0.192*** (0.060)
Laos	0.079* (0.042)	5.134*** (0.890)	0.389*** (0.035)	0.029 (0.043)
Malawi	-0.094*** (0.018)	5.226*** (0.456)	0.391*** (0.016)	0.274*** (0.017)
Mexico	-0.181*** (0.004)	6.287*** (0.084)	0.221*** (0.003)	0.084*** (0.003)
Nicaragua	-0.284*** (0.017)	1.495*** (0.395)	0.090*** (0.013)	0.477*** (0.041)
Niger	-0.181*** (0.028)	6.537*** (0.773)	0.498*** (0.027)	0.254*** (0.032)

*Continued*

**Table 3** Demographic differences of public sector—cont'd

	(1)	(2)	(3)	(4)
	Male	Age	Secondary or tertiary education	Tertiary education
Nigeria	-0.015 (0.029)	8.820*** (0.749)	0.147*** (0.027)	0.391*** (0.027)
Pakistan	-0.022*** (0.006)	7.256*** (0.196)	0.285*** (0.008)	0.162*** (0.007)
Panama	-0.153*** (0.014)	6.708*** (0.330)	0.178*** (0.010)	0.300*** (0.014)
Peru	-0.115*** (0.025)	9.617*** (0.566)	0.027*** (0.009)	0.330*** (0.022)
Serbia	-0.055*** (0.013)	-1.147*** (0.311)	0.248*** (0.010)	0.226*** (0.011)
South Africa	-0.175*** (0.005)	5.184*** (0.101)	0.041*** (0.003)	0.138*** (0.004)
Sri Lanka	-0.216*** (0.039)	2.624*** (0.823)	0.165*** (0.020)	0.326*** (0.035)
Tajikistan	-0.067** (0.026)	7.794*** (0.680)	0.145*** (0.018)	0.361*** (0.025)
Tanzania	-0.039 (0.030)	8.792*** (0.748)	0.536*** (0.028)	0.118*** (0.020)
Timor Leste	-0.056** (0.028)	2.658*** (0.746)	0.323*** (0.035)	0.044** (0.018)
Uganda	-0.020 (0.041)	9.086*** (0.934)	0.388*** (0.038)	0.360*** (0.042)
United Kingdom	-0.224*** (0.005)	3.787*** (0.137)	0.126*** (0.004)	0.253*** (0.005)
United States	-0.085*** (0.012)	4.314*** (0.302)	0.020*** (0.002)	0.159*** (0.010)
Vietnam	-0.023 (0.031)	5.589*** (0.660)	0.175*** (0.018)	0.305*** (0.030)

Standard errors in parentheses.

Each table cell contains a within-country difference of means of demographic characteristics of public sector and private sector employees. Ambiguous sectors such as NGO are omitted. Each country's data come from a single household-level survey, as listed in the [Appendix](#).

\*,  $p < 0.10$ ; \*\*  $p < 0.05$ ; \*\*\*  $p < 0.01$ .

motivated than private sector workers on average, this difference depends on the corruption level of the country. In countries with high levels of corruption, intrinsically motivated individuals are not more likely to join the public sector. Two recent laboratory studies are consistent with this association. [Hanna and Wang \(2014\)](#) have students from a university in India play a series of experimental games designed to measure various personality traits, such as cheating and prosocial behavior. The authors find that students

who cheated in a random dice game are more likely to express interest in a public sector job. In a corruption experiment with private sector job aspirants and aspirants of the Indian bureaucracy, [Banerjee et al. \(2015\)](#) examine embezzlement of resources in which “supervisors” evaluate the performance of “workers” and then pay them. They find that aspirant bureaucrats are more corrupt than private sector aspirants, but their likelihood of being corrupt is similar across sectors.

This tradeoff in vocational profiles has also received theoretical attention regarding how best to design personnel policy. In addition to the intrinsically motivated, governments and other mission-driven organizations value individuals of high quality as well. But if higher quality candidates demand more compensation, then higher wages may be needed to attract these individuals. But as a small theoretical literature in economics [e.g., [Delfgaauw and Dur \(2007\)](#), [Francois \(2000\)](#), and [Prendergast \(2007\)](#)] has pointed out, offering higher wages may come at the cost of attracting individuals who are more corruptible or care less about the mission. Whether this tradeoff exists is ultimately an empirical question, which has recently led several scholars to explore the extent to which financial incentives can affect a government’s ability to recruit publicly motivated and high-quality individuals. In the following section, we review a nascent experimental literature on how certain job attributes, including compensation, affect two aspects of the recruitment process into the public sector: who applies for the job and who accepts the job. We then conclude with a short discussion of the empirical evidence on whether intrinsic motivation does in fact lead to higher job performance in the public sector.

### 3.1 Financial incentives

#### 3.1.1 Effects of financial incentives on the applicant pool

If higher quality candidates, as priced by the market, demand higher compensation, then higher wages in the public sector are necessary to attract those candidates. Does offering higher wages come with the cost of attracting candidates with weaker public service motivation? This question has motivated three recent experimental studies. [Dal Bo et al. \(2013\)](#) implemented a field experiment as part of an official program of Mexico’s federal government called the Regional Development Program (RDP). The program, which sought to enhance state presence in 167 of Mexico’s most marginalized municipalities, conducted a recruitment drive to hire 350 community development agents who were tasked with the responsibility of identifying areas where public good provision is deficient and working with existing public programs and local authorities to remedy such deficiencies.

A unique feature of this recruitment drive was the exogenous assignment of wage offers across recruitment sites. Two different wage offers were randomly assigned across 106 recruitment sites. In one set of recruitment sites, the program offered 5000 pesos per month, while in the other sites, the program offered a wage of 3750 pesos. Candidates who were interested in this position were then required to undertake a screening exam that was designed to measure various dimensions of quality and motivation.

[Dal Bó et al. \(2013\)](#) find that higher wages do help attract a higher quality candidate pool. In the places that announced a higher salary, the average applicant was smarter, had better personality traits, had higher earnings, and had a better occupational profile (e.g., more experience and white collar background). Moreover, contrary to theoretical concerns, these effects do not come at the cost of attracting less publicly motivated candidates, as measured by their performance on a public service motivation inventory.<sup>7</sup>

An important design feature of this study was its ability to offer two different wages for the exact same position, which was in large part due to the size of the program and its expansive geographical coverage. In other settings, where offering different wages has not been feasible, researchers have had to adopt alternative, creative approaches to addressing this question. One such example is a study by [Ashraf et al. \(2015\)](#). In this study, the authors partnered with the Government of Zambia to hire approximately 330 community health care workers. Instead of offering different wages, the authors introduced experimental variation in the how the position was advertised. In 24 of the 48 districts, potential candidates saw a job advertisement that highlighted the job's promotion prospects and the opportunity for career advancement. In the other districts, applicants saw a poster that emphasized the social importance of the job.

This recruitment process led to over 2400 applicants. In the districts where the job ads stressed career incentives, applicants were much more qualified as measured by their high-school test scores and past performance in their natural science courses. These applicants also displayed a high degree of prosocial motivation, with levels that were similar to the applicants that applied under the social incentive treatment. While the applicants who applied under the career incentives treatment did place a higher weight on career benefits, the authors conclude that making career versus social incentives salient did not induce a tradeoff between a higher quality applicant pool and a prosocially motivated one.

In contrast to these two studies, [Deserranno \(2015\)](#) finds that financial incentives can lead to a less socially motivated applicant pool. Her field experiment was conducted in rural villages of Uganda in collaboration with the NGO, BRAC. The recruitment drive was for health promoters, which was a position that did not previously exist and whose remuneration was uncertain since it depended on the sales of health products. The experiment exploits these two features of the position to introduce variation in how the financial aspects of the job were advertised. In one treatment arm, the job advertisement

<sup>7</sup> Perry's Public Service Motivation Index ([Perry, 1996](#)) is the most commonly used measure of intrinsic motivation in the public sector. This index is constructed based on a questionnaire in which the subject must express agreement or disagreement with each of 32 statements. The questionnaire elicits opinions on the attractiveness of politics, public service, and prosocial activities. The questionnaire is subdivided into six modules labeled "Attraction to Policy Making," "Commitment to Public Interest," "Social Justice," "Civic Duty," "Compassion," and "Self-Sacrifice". Each dimension is an average of responses to several statements that are measured on a Five-point Likert scale, where a five represents strong agreement with the statement, and a one denotes strong disagreement.

mentioned the minimum amount that health promoter was expected to earn (low-pay treatment) and, in another treatment arm, advertised the maximum amount a health promoter was expected to earn (high-pay treatment). A third treatment advertised the mean of the expected earnings distribution (medium-pay treatment).

The study finds that while the high-pay treatment attracted 30 percent more applicants relative to the low-pay treatment, the applicants had much less experience as health volunteers and were much more likely to state “earning money” as the most important feature of the job. Applicants under the medium- and high-pay treatments were also 24 percentage points less likely to make a donation to a public health NGO in the context of a dictator game. Although the author finds large effects on these various measures of intrinsic motivation, she does not find treatment effects on candidate quality, as measured by the applicant’s education and income.

In sum, the conclusions from a scant experimental literature are mixed, but this is not at all surprising. Putting aside differences in the actual treatments, as [Dal Bo et al. \(2013\)](#) point out in their model, whether financial incentives crowd out the intrinsically motivated will depend on how these personal traits (e.g., intrinsic motivation versus quality) are correlated within the broader population—correlations which the literature has yet to document in a systematic way. These three studies, which were conducted in very different contexts, represent a step forward in the literature, but without more information on how personality traits vary across broader populations, it is difficult to make general conclusions about the exact tradeoffs that financial incentives induce.

### ***3.1.2 Effects of financial incentives on recruitment***

The power of wages is not limited to attracting a larger and better applicant pool. Higher wages also increase an organization’s ability to fill vacancies. In the [Dal Bo et al. \(2013\)](#) study, the authors found that the Mexican government was 35.2 percent more likely to fill the vacancy when offering the higher wage, which corresponds to a short-run labor supply elasticity of 2.15. This elasticity is similar to other quasi-experimental estimates found in the literature (e.g., [Manning, 2011](#)), as well as one reported in [Deserranno \(2015\)](#). Even though the study by [Deserranno \(2015\)](#) manipulates earning expectations (as opposed to actual earnings), the author finds an experimental elasticity of 1.8 when comparing takeup in the low-pay treatment group to takeup in the high-pay treatment group.

Part of the reason why higher wages lead to higher recruitment rates is because they help to compensate for aspects of the job that a candidate dislikes. This mechanism was on clear display in the study by [Dal Bo et al. \(2013\)](#). Although the applicants for the RDP position were all applying for the same job, the jobs were located in different municipalities throughout the country. At the time of the application, the candidates did not know where the job was located and were only told this information during the offer stage. As a result, jobs that were ex-ante quite similar became quite different ex-post depending on

where the job was located and the characteristics of the municipality. The authors show that distance to the municipality (from their current residence) and attributes such as the level of drug violence and the lack of public goods in the municipality were all important hurdles to filling the vacancies. Fortunately, however, higher wages proved to be an effective instrument in clearing these hurdles.

### 3.2 How should governments screen?

Wage offers affect who applies for government jobs, but government jobs are typically oversubscribed, so government—like all employers—needs to winnow down the set of applicants to those they hire. Governments vary in the way they screen their public servants. Some rely on the passage of civil service exams or attainment of university degrees, while others adopt more discretionary approaches that, while permitting them more flexibility, can also be prone to corruption and patronage. These different screening strategies have important implications for not only the quality and performance of the bureaucracy but also for the type of person who applies.

An important consideration for any organization when screening and selecting personnel is match quality. If employers and employees share a common vision and objectives, then this positive match quality increases organizational efficiency and diminishes the need for high-powered incentives. For governments, who are responsible for providing public goods that are difficult to price in the market, the ability to recruit public service—motivated individuals might be especially beneficial ([Besley and Ghatak, 2005](#)).<sup>8</sup>

Arguably, the recruitment of publicly motivated individuals has other benefits as well. Individuals with high levels of intrinsic motivation are less likely to shirk in an environment where incentives are low powered and/or when noncontractible elements of the service provision exist ([Francois, 2000](#)).

In support of these theoretical arguments, a large empirical literature in public administration shows that intrinsic motivation—and specifically public service motivation—is associated with higher levels of performance in government work ([Perry and Hondeghem, 2008](#)).<sup>9</sup> Recently, economists have begun to contribute to this literature.

<sup>8</sup> Before we can answer the question whether or not government should screen on intrinsic motivation, we must take a step back and ask whether it even measurable and quantifiable. There is a rich and growing literature in psychology and economics that suggest that personality traits including intrinsic motivation can be measured (see, e.g., [Almlund et al., 2011](#)).

<sup>9</sup> In the public administration literature, recent meta-studies suggest that public-service motivation is positively correlated with job performance in the public sector, broadly defined ([Petrovsky, 2009](#)). [Naff and Crum \(1999\)](#) use a sample of over 8000 U.S. federal employees and find that public-sector motivation correlates with individuals' last performance evaluations. [Park and Rainey \(2008\)](#) analyze data from 22 federal agencies in the United States and find that public service motivation is positively correlated with self-reported measures of job productivity and quality of work. Similar results are found using government data from Switzerland ([Ritz, 2009](#)) and the Netherlands ([Steijn, 2008](#)).

For example, as part of a monitoring experiment of health clinics in the district of Punjab, [Callen et al. \(2015\)](#) examine the job performance of clinic doctors. They find that those who scored higher on the public service motivation index are much less likely to shirk and falsify health reports. In the same study discussed previously, [Deserranno \(2015\)](#) also finds, in the case of the Ugandan health promoters, that prosocial motivation is a strong predictor of job performance. Health promoters who had donated a greater share of their endowment to a local NGO visited a larger number of households, provided more prenatal checks, and organized more public presentations. [Dizon-Ross et al. \(2015\)](#) conduct a survey on nurses of antenatal care centers in Uganda, Ghana, and Kenya as part of an audit study on bednet distribution programs. They find that nurses not only exhibit high levels of prosocial motivation but that it is predictive of job performance.

Despite the mounting evidence linking public service motivation to job performance, establishing causality has proven challenging. Given the difficulties in directly generating experimental variation in a person's level of intrinsic motivation, researchers have had to rely on indirect approaches. One approach has been to introduce experimental variation in who applies for the same job. The experiment induces a selection effect, while keeping any potential incentive effect constant. Both the studies by [Ashraf et al. \(2015\)](#) and [Deserranno \(2015\)](#) provide examples of this approach. In the study by [Ashraf et al. \(2015\)](#), the authors used the two different recruitment strategies to create variation in the type of health promoters that were recruited. Once employed, all the health workers were tasked with the same responsibilities and faced the same incentives. Based on this design, they find that health workers attracted by career incentives are much more effective at delivering health services, as measured by home visits and the organization of community meetings. These health promoters were also more likely to remain in their posts over the course of 18 months. Although these results imply a negative relationship, if any, between intrinsic motivation and performance, it is worth noting that the level of prosocialness among the health promoters who were recruited in the career incentives treatment was also quite high.

The study by [Deserranno \(2015\)](#) provides stronger evidence in support of the relationship between prosocialness and job performance. Among the agents recruited under the low-pay treatment who were measured to be more prosocial, she finds higher aggregate performance in the first year of work. Compared to the high-pay treatment, the health promoters recruited under the low-pay treatment visited a larger number of households, organized more public presentations in the village, and provided more natal checks. She also finds that they were more likely to target the most vulnerable households.

While these studies can credibly identify the effects of "selection" on job performance, what this selection effect comprises is not entirely clear. Prosocial motivation is frequently found to be correlated with various other personal traits, including the Big Five. Short of randomly assigning individuals based on a specific attribute, it is difficult to separate the effects of intrinsic motivation from other positive personality traits.

Another experimental approach that studies have explored has been to test whether a particular intervention is more effective among individuals with high levels of intrinsic motivation. For example, the goal of the experiment in the study by [Callen et al. \(2015\)](#) was to reduce high levels of absenteeism among clinic doctors and staff in Punjab.<sup>10</sup> One source of this absenteeism was the fact that these clinics were rarely inspected, and when inspections did occur, doctors and inspectors would collude and falsify the report. In collaboration with senior health officials of the Department of Health, the authors introduced a new monitoring program in 18 of the 35 districts constituting their experimental sample. In the treatment districts, the traditional paper-based monitoring system for clinic utilization and worker absence was replaced with a smartphone application. The new system allowed health system inspectors to upload the results of their assigned visit to a central dashboard which instantly updated reports at different levels of aggregation. The data, which included geo-tagged, time-stamped facility staff photos, made it difficult for the inspector to falsify his report. While the study finds that the monitoring technology did increase the number of inspections, there was significant heterogeneity in the treatment effects by the personality type of the inspector. Higher quality inspectors responded much more positively to the treatment. A Big Five index one standard deviation higher, for example, is associated with a differential 35 percentage point treatment effect in terms of health inspections.

In another example, [Bellé \(2012\)](#) conducted two experiments with nurses at a large public hospital in Italy. He was interested in understanding whether public service motivation interacts with two interventions that social psychologists have found to be effective at stimulating job performance: (1) beneficiary contact ([Grant et al., 2007](#)) and (2) self-persuasion interventions ([Aronson, 1999](#)). The first intervention is based on the idea that contact with customers, clients, and beneficiaries outside of the organization can help to motivate employees to perform more effectively. The second experiment is based on the premise that employees are most likely to be influenced by credible and trustworthy sources, and one such source is the person himself. Based on this theory, researchers have shown that employees find public service more important after they were asked to reflect on the importance of public service and then made to publicly advocate for it both in writing and in person.

The study by [Bellé \(2012\)](#) was based on a sample of 90 nurses, randomly assigned across the two treatment interventions and a control. The nurses were tasked with assembling surgical kits that were being shipped to a former war zone that was facing a humanitarian emergency. Based on this task, the study examined three principal performance measures: (1) the number of minutes each participant contributed to task, (2) the number

<sup>10</sup> Based on unannounced visits to clinics at baseline, they found that only 56 percent of clinics had been inspected in the prior 2 months and that 32 percent of clinics had no doctor present.

of surgical kits that each participant assembled during her shift, and (3) the average number of surgical kits each participated completed per minute. The author finds that the effects of both interventions were stronger for employees who had a higher prosocial motivation level at baseline.

While these studies establish that the characteristics of individuals are an important determinant of performance and they suggest how governments could change the applicant pool, they do not necessarily tell us how governments should screen among the candidates who apply to further improve selection. [Hanna and Wang \(2014\)](#), for example, show that current Indian civil service–type screening exams would not eliminate the negative selection on dishonesty they find in their setting. Given the large public sector premia we observe for low-income countries, an important friction in improving the human capabilities of the state may lie in its screening technologies. To understand how to most effectively screen seems an important direction for future research.

#### 4. USING INCENTIVES TO IMPROVE PERFORMANCE

Once selected, it may be possible to use incentives to further improve workers' performance. Public sectors careers, however, typically feature a relatively flat incentive structure. As discussed previously, in the standard civil service model, adopted almost universally since the early 20th century to limit politician discretion over appointments and salaries, the public sector is staffed by salaried civil servants whose salaries are based on rigid and formulaic pay scales. These pay scales feature compressed wages relative to that in the private sector [e.g., [Borjas \(2015\)](#) for the United States]. The combination of formulaic pay systems and wage compression limits the degree to which financial incentives can be used to reward the performance of public servants: the formulas are largely based on seniority and position, allowing little room for discretion, and the wage compression is such that even promotions within the civil service are less of an incentive than in the private sector.

While this type of salary structure may be appropriate for governments in many contexts, it may not be effective in developing countries where government officials are often thought to have poor job performance. Therefore, scholars have begun to examine the costs and benefits of providing additional incentives to government workers in developing countries. In this section, we review recent experimental evidence that seeks to shed light on these issues in a variety of sectors. We begin first with evidence on using financial incentives to reward good performance. The evidence not only sheds light on the degree to which such incentives can improve performance but also highlights the challenges with using such incentives in practice, particularly those that arise when using them in the public sector.

Given the constraints in financial payments imposed by civil service systems, we then go on to consider nonfinancial incentives, which are prevalent in government contexts.

We examine one type, in particular, that is quite common in government practice: using transfers to more or less desirable postings as an incentive device. We then examine other types of nonfinancial incentives.

## 4.1 Financial incentives

Government officials do many types of jobs, and some are easier to incentivize than others. In some cases, what we refer to as “agents of government authority” are those who are tasked with ensuring citizens comply with government laws and regulations. For such officials—such as police, judges, prosecutors, tax inspectors, building inspectors, and so on—there is a natural tension between what the government would like the agent to do (e.g., to make people pay taxes that are due under the law) and what the targets of government enforcement would like the agent to do (e.g., to allow them to avoid paying taxes). This tension invites opportunities for corruption between the agent and the citizen (e.g., reducing taxes in exchange for a bribe) and, as we will see, complicates the incentive problem for the government. In other cases, such as frontline government service providers, the government and the citizen’s incentives are aligned: both would like the agent (e.g., the teacher) to provide more or better services. Providing incentives may therefore be more straightforward in the second case. We consider incentives in both contexts in turn.

### 4.1.1 *Incentives for agents of government authority*

#### 4.1.1.1 Incentives for tax collection

Several recent experiments explore the risks and rewards of financial incentives in the public service. [Khan et al. \(2014\)](#) conduct a field experiment in urban Punjab, Pakistan, to study performance pay for tax inspectors. The experiment involved high-powered financial incentives for property tax inspectors who are in charge of assessing properties, collecting property taxes, and levying sanctions on those who fail to pay. The basic treatment gave the team of tax staff in an area, which consisted of three people, an incentive payment equal to an average of 30 percent of tax revenues collected above a historically predicted benchmark, enough to double their baseline wages. Tax inspectors are exactly the sort of government worker where one might be concerned ex-ante about the efficacy of incentives; while there is substantial scope for improvement through either increased effort or reduced corruption, incentives also have the potential to increase bribes by raising the bargaining power of tax inspectors (who now must be paid a higher bribe to compensate them for their foregone incentive payment) or to lead to overtaxation as was thought to be the case historically.

[Khan et al. \(2014\)](#) find evidence for both the positive and negative aspects of incentives. On the plus side, the incentives raised revenue substantially. On average, treated areas had revenue growth that was 9.3 log points greater than control, which translates to a 46-percent higher growth rate in revenue. Incentive schemes that only rewarded

on revenue did best, increasing revenue growth by 12.8 log points (62-percent higher growth), whereas those that attempted to control multitasking problems through incentive schemes that also rewarded taxpayer satisfaction and accuracy of tax assessments had less impact on revenue yet did not improve these dimensions. The revenue gains substantially exceeded the costs of the incentives. The incentives did not appear to reduce taxpayer satisfaction, in part, because the increased tax collection was concentrated among a small number of taxpayers. On the negative side, however, they find evidence that bribe rates did increase in incentive areas, potentially to compensate incentivized tax inspectors for foregone incentive payments.<sup>11</sup>

#### **4.1.1.2 Incentives for policing and justice**

The study by Khan et al. (2014) is relatively rare in focusing on the tax sector, where the government has a potentially adversarial role against the taxpayer, which leads to opportunities for collusion and where incentives can have perverse effects, such as the increase in bribes they document. The police force is another area with agency problems, where one might be concerned that financial incentives (e.g., on the number of citations issued, arrests made, or the like) could lead to overzealous or inaccurate enforcement or simply a reallocation of resources from nonincentivized to incentivized tasks. Baicker and Jacobson (2007), for example, document that when police agencies in the United States are allowed to keep the revenue they obtain from assets they seize in drug arrests, they increase drug arrests but do so by reducing enforcement of other petty crimes, suggesting that multitasking is an important issue. A commonly voiced concern about these laws that their paper does not address is whether these type of laws lead to unjustified seizures and abuses (Miller and Selva, 1994). Other similar areas where financial incentives have been tried, but not rigorously evaluated, include incentives for prosecutors in New York City to ensure speedy disposition of cases (Church and Heumann, 1989), as well as historical examples from the United States where prosecutors were paid incentives based on conviction rates (Meares, 1995). Exploring the impact of incentives in these areas in a more rigorous and careful way, and seeing whether they are effective in developing countries with more corrupt and generally less-effective police forces, seems an important area for future work.

#### **4.1.2 Incentives for frontline service providers**

A more common area of focus has been incentives for health and education service providers. These have taken two broad forms, incentives on outcomes (e.g., test scores, immunizations given) and incentives on inputs (e.g., provider attendance).

<sup>11</sup> Other nonexperimental studies of tax agencies also find increases in revenue (e.g., Kahn et al., 2001; Burgess et al., 2010) but are unable to examine the potential downside in terms of over-enforcement or bribery.

#### 4.1.2.1 Incentives on outcomes: test scores

One commonly considered type of outcome-based incentive is teacher incentives based on student test scores.<sup>12</sup> [Muralidharan and Sundararaman \(2011\)](#) report the results of a large-scale teacher incentive program run by the Indian state of Andhra Pradesh as a school-level randomized trial. Public school teachers were paid incentives based on test scores, with both group and individual incentives considered. Incentives were substantially smaller than in the study by [Khan et al. \(2014\)](#) discussed previously, as they were calibrated to be around three percent of a typical teacher's annual salary. They find that the incentives were effective in promoting learning: after 2 years, students in incentivized schools had test scores that were 0.27 standard deviations higher in math and 0.17 standard deviations higher in language. They find no evidence of multitasking; in fact, students also do better in nonincentivized subjects, such as science and social studies. Incentives appear to have worked by increasing effort conditional on attendance not by increasing teacher attendance. The individual incentives outperformed the group incentives by the end of the second year. [Muralidharan \(2012\)](#) reports that the effects increase even more with time: after 5 years, students in treatment schools had test scores 0.54 standard deviations higher in math and 0.35 standard deviations higher in language and still had higher test scores in nonincentivized subjects.

On the other hand, [Glewwe et al. \(2010\)](#) find somewhat less encouraging results. They conduct a randomized trial of teacher incentives in Kenya, where an NGO provided in-kind prizes to teachers in Kenyan government schools on the basis of school-level performance on district exams, where those who did not take the exam were imputed a low score. They find that incentivized schools had more people taking the government exam and higher scores on the government exam used for the incentives. However, unlike the example by [Muralidharan and Sundararaman \(2011\)](#) that found positive spillovers to nonincentivized subjects, they find no evidence of higher scores on an independent exam administered by the NGO that was not linked to performance incentives. The authors conclude that multitasking was a real issue in their context and that teachers may have emphasized test-taking skills, as opposed to general instruction, in response to the incentives. In both cases, incentives improved targeted indicators, but understanding why there were positive spillovers to nonincentivized contexts in India but not in Kenya seem an important area for future research.

#### 4.1.2.2 Incentives on outcomes: health

While incentives based directly on health outcomes are rare, one notable example is the study by [Miller et al. \(2012\)](#). [Miller et al. \(2012\)](#) conducted a randomized trial in 72 Chinese primary schools in which school principals received performance payments based on

<sup>12</sup> In addition to the experimental studies reviewed here, there is also a large literature on teacher performance pay in the United States. See [Neal \(2011\)](#) for a review.

reduction in anemia among their students. Specifically, principals were paid 150 RMB per student who changed from anemic to nonanemic over the course of the intervention. This implied a payment of roughly 2 months' salary for reducing anemia by half. Comparison groups were given the same information and subsidies as the incentive treatment, but no direct financial incentives. They find that the incentives reduced anemia compared with the pure control group by about 5 percentage points (23 percent). The nonincentivized comparison groups did not achieve statistically detectable reductions in anemia, suggesting a role for incentives, but confidence intervals are such that they cannot statistically distinguish between the incentive group and the nonincentive information and subsidy groups. They report that incentivized school principals were more likely to use subsidies for iron-focused supplements, whereas nonincentivized school principals used subsidies for supplements that could affect both iron and overall calorie intake. Depending on one's perspective, this could be considered a multitasking issue as well to the extent that one is interested in both types of supplementation.

#### 4.1.2.3 Incentives on service delivery

In health, incentives for providers have tended to be focused instead on measures of service delivery, such as the number of immunizations given. These service delivery metrics can be thought of as somewhere between ultimate outcomes (e.g., learning, lack of disease) and provider inputs (e.g., attendance). One reason for focusing on this level is that for ultimate health outcomes, the signal to noise ratio may be high; that is, in a given context, most of the variance in health outcomes is idiosyncratic rather than due to provider effort. If one believes there is a clear mapping from health service delivery to health, these types of incentives may make sense.

[Basinga et al. \(2011\)](#) and [Gertler and Vermeersch \(2012\)](#) examine this approach. The intervention they study in their experiment took place in Rwanda and provided incentives to primary care facilities, which were in turn used to compensate facility personnel. The incentives were based on the quantity of visits to the facility for various services (e.g., childbirth in the facility, prenatal care) and the content of services provided in those visits (e.g., pregnant women receiving tetanus vaccines and malaria prophylaxis during prenatal care, immunizations given during postnatal care, etc.), weighted by an overall quality index of the facility. The incentive payments were substantial, equal to a 38-percent increase in total compensation for facility personnel. They found that the incentives led to a substantial increase in prenatal and postnatal services, which translated into increased health: infant weight-for-age increased by 0.53 standard deviations and height-for-age for children aged two to five by 0.25 standard deviations, with increased breastfeeding and reductions in infant illness hypothesized to be important channels.

[Olken et al. \(2014\)](#) report the results of a large-scale field experiment in Indonesia in which villages were provided with incentive payments based on health service delivery (similar to that in Rwanda), school enrollment, and education. Specifically, villages

received a block grant each year that they could use for any purpose related to health or education. In incentivized areas, 20 percent of the total amount set aside for block grants in a subdistrict was allocated to villages based on their performance on the targeted health and education indicators; in nonincentivized areas, the block grant was allocated based only on population. The incentive was to the community as a whole and, unlike the previous examples, was generally not passed on to service providers but was instead used for programming (e.g., nutritional supplements, subsidies for childbirth, etc.). Comparing the incentivized to nonincentivized areas, they found the incentivized areas performed better on the targeted health indicators. On average, the eight targeted health indicators were about 0.04 standard deviations higher in the incentivized than nonincentivized areas. These effects were about twice as large in areas with low initial levels of performance, but the relative gain of incentivized to nonincentivized areas declined over time as nonincentivized areas improved. The main health reduction was a 15-percent (2.6 percentage points) decline in malnutrition rates, though again this effect became more muted over time. There were no detectable differences between incentivized and nonincentivized areas on educational outcomes. The program suggests that the incentives sped up improvements on the targeted health outcomes, with no detectable multitasking effects.

#### 4.1.2.4 Incentives on inputs: provider attendance

The final category of financial incentives that we consider is incentives based on attendance. Given the problems with provider attendance highlighted by Chaudhury et al. (2006), this is clearly an important issue—but key questions are whether attendance responds to financial incentives and, if so, if it translates into ultimate outcomes. Here the evidence is mixed. In an experimental study for teacher attendance at single-teacher schools run by NGOs in India, Duflo et al. (2012) provided linear incentives based on the number of days (above 10 per month) that teachers could submit time-stamped photos of themselves with students to prove they had been attending. They found that the incentives not only increased attendance but also led to increased learning, with students in schools where teachers were incentivized to attend having test scores about 0.17 standard deviations higher than in control schools after 1 year.

On the other hand, a similar study by Banerjee et al. (2008) of financial incentives for nurses' attendance in India provides a more cautionary note. In that experiment, the incentives for attendance were broadly similar to those in Duflo et al. (2012): nurses who were recorded absent more than 50 percent of the days in a month would have their pay reduced by the number of days they were recorded absent, and nurses who were absent more than 50 percent of the days in two consecutive months would be suspended from government service. Nurses used a protected time/date stamp machine to verify attendance. In their study, while there was initially a substantial treatment effect, the effect diminished over time and was zero at the end of their study. Although they do not

have the data to confirm this, anecdotal evidence suggests the decline was due to nurses learning how to exploit loopholes in the systems and recording more exempt absences over time. One possible difference is that the [Duflo et al. \(2012\)](#) schools were run by an NGO, which may have had more independence in enforcing the incentives than the government. A key question then is understanding which of these effects is more likely to generalize: the positive long-run effects found in the study by [Duflo et al. \(2012\)](#) or the rapid decline in effectiveness found in the study by [Banerjee et al. \(2008\)](#).

## 4.2 Nonfinancial incentives

While the majority of work has focused on financial incentives, nonpecuniary incentives are potentially important. While civil service regimes typically place many restrictions on hiring and firing, they have much more flexibility in assigning bureaucrats to postings within the civil service and these postings can be used as reward and punishment devices. Many bureaucracies informally recognize high achievers (i.e., “employee of the month” type awards). Public sector jobs, in particular, may seek to take advantage of the fact that their employees may be public spirited and use this as a way of creating rewards. While much less extensively studied than pecuniary incentives, several studies suggest that it is a promising direction for further exploration.

### 4.2.1 Transfers and postings

Civil service regimes typically feature much more flexibility in where people are posted than in whether people are fired or how much they are paid. This is perhaps natural, in that there are a wide variety of positions that need to be filled and these positions are heterogeneous in many dimensions, both in terms of the skills needed to complete the job effectively and their desirability as a place to work. In both cases, there can be a mix of common and idiosyncratic rankings. For example, as we discussed in the case of Mexico RDP program, civil servants were much more willing to work in a safe community, as opposed to one with high incidences of drug violence; this would be a common preference. The civil servants also preferred to work near their place of residence; since people are from different places, this creates idiosyncratic preferences. The same can be said for job attributes: a common attribute would be the need to put the cleverest tax inspectors on the most complicated corporate tax cases; an idiosyncratic attribute would be the need to match police with areas where they have social connections they can use to gather information. To the extent that there are common components to preferences, this creates scarcity for posting in the most desirable locations, and such plum postings can be used as an incentive device.

One problem with transfers as an incentive device is that politics often gets in the way. [Iyer and Mani \(2012\)](#) examined a comprehensive data set that tracks the careers of elite Indian Administrative Service personnel. They show that transfers are likely right after a new Chief Minister is elected, particularly for those bureaucrats who were not at the very

top of their initial class in terms of performance. However, even though ability is predictive of future success, caste affinity to the politician also plays an important role. Bureaucrats who share the same caste as the chief minister's party are just as likely to be assigned to important posts as the high-ability bureaucrats. These results, while not definitive, suggest that while transfers are quite common, they are not entirely based on performance, which may dampen their usefulness as a performance tool.

[Banerjee et al. \(2014\)](#) explore several aspects of these issues in the context of the police force in Rajasthan, India, a context in which transfers are frequent: one-third of all policemen were transferred during a typical 18-month period. As in the elite civil servants examined by [Iyer and Mani \(2012\)](#), anecdotal evidence suggests that police transfers are frequently imposed by politicians, often for reasons that may reduce their use as incentive device (e.g., for partisan or corrupt motives). They explore a treatment where all transfers were frozen during a 2-year period, except for well-documented cases of police misconduct. The idea was to remove arbitrary transfers and leave only transfers being used as an incentive device. They find that this freeze had no effect on outcomes such as whether decoy surveyors were treated differently or community satisfaction. One potential reason for the lack of effect is that, if the exceptions were sufficiently difficult to implement, the freeze could have eliminated both transfers used as incentives as well as politically motivated transfers. The elimination of the transfer-as-incentive could have then offset the positive effect of removing political transfers.

Consistent with this idea, a second treatment suggests that transfers have the potential to be used as an incentive device. In a second experiment, [Banerjee et al. \(2014\)](#) examined an anti-drunk-driving reform, where police were supposed to conduct sobriety effects. Two groups of police ran sobriety checks. In the first group, they worked with police in the central reserve "police lines" group, who are outside of typical station assignments and was given the incentive that they would be transferred back to the regular police unit if they performed well. The second group consisted of police from normal stations, who had no transfer incentives. They randomized which units were sent to which areas and found that those in the first group performed better in terms of whether the roadblock actually occurred, the number of people stopped, and so on. Of course, the composition of personnel was different in the two groups, so one cannot know if it is the transfer incentive per se that is driving the results or some other factor (e.g., maybe the police line teams had nothing else to do with their time, while the regular police teams were juggling many other tasks), but the results are suggestive that this could be important.

#### 4.2.2 *Intrinsic motivation*

Beyond explicit incentives, it may be possible to use other types of intrinsic rewards as a motivational tool. One experimental study that examines this idea is the study by [Ashraf](#)

[et al. \(2014\)](#). In their study, public health extension workers who are tasked with selling condoms are randomly assigned to either different financial rewards (margins of 10 percent or 90 percent on each condom sale) or a nonfinancial reward that gives agents additional stars for each sale on a thermometer-type display. They find remarkable evidence of the effectiveness of nonfinancial rewards: the thermometer treatment agents sell twice as much as those in the financial rewards treatment.

Another recent, nonrandomized study also suggests that dimensions other than incentives may be important to job performance. [Rasul and Rogger \(2015\)](#) use a survey to measure management practices in the Nigerian bureaucracy and find that autonomy is positively correlated with job performance, whereas performance incentives are negatively correlated. Of course, there could be endogeneity problems: one might choose to give performance incentives to those bureaucrats who behave badly and to reward high performers with autonomy.

### 4.3 Summary

Several themes emerge from the evidence reviewed in this section. First, there is robust evidence that financial incentives matter: across a wide variety of settings, financial incentives in a government context seem to increase performance on the incentivized dimension. This is not particularly surprising. In fact, given these robust results, the question becomes why do governments not use financial incentives more often? Part of the reason can be attributed to the simple fact that unlike in the private sector where firms can contract on profits, performance in the public sector can be hard to measure. Also, as we discussed, there is some evidence that multitasking issues can be a problem—yet we do not have a clear understanding of when multitasking issues will or will not be present. Finally, as [Benabou and Tirole \(2006\)](#) have highlighted, financial incentives may even reduce effort among the prosocially motivated. While a number of studies have documented such adverse effects in such activities as volunteer work (e.g., [Gneezy and Rustichini, 2000](#)) or blood donation (e.g., [Mellstrom and Johannesson, 2008](#)), it remains to be seen whether this is a first-order issue in the context of governments. Developing a clearer understanding of when these issues will be present seems an important direction for future research.

Finally, research on financial incentives has primarily focused on front-line service providers, where the agent's principal (e.g., bureaucrats at the central government ministry) and the citizens served by the front-line agent have aligned incentives. With just a few exceptions, there has been much less work on the more complex case where these interests are unaligned, such as tax, police, procurement, and so on. Understanding the degree to which incentives can be effective in this context without further empowering these officials to collect more bribes or overenforce the law seems an important area for future work.

## 5. MONITORING MECHANISMS AND PUBLIC SERVICE DELIVERY

### 5.1 Overview

Incentives focus on tying rewards—typically financial rewards—to easily observable and verifiable measures of performance. Taxes is a canonical example: the state easily observes the amount of taxes collected by each tax inspector and so can base rewards on that. In many cases, however, monitoring performance itself requires costly effort on the part of either state or nonstate actors. We now turn to whether improved monitoring can improve the performance of civil servants.

Increased monitoring can improve program performance via multiple channels. First, in cases where outcomes are not observed without some effort, increased monitoring can allow managers to directly enforce punishments and rewards based on program outcomes (e.g., firing or transferring poor performers). Second, monitoring can play an important deterrence role. Third, access to monitoring results can empower citizens to demand and obtain better services by threatening to report on or vote out poor performers.

However, there are also reasons to believe that information alone may not suffice. First, in situations where state capacity is weak, managers' or regulators' ability to impose punishments is limited, and improving information flows may do relatively little by itself. Second, those in charge of collecting information or monitoring based on available information may themselves be susceptible to corruption and misuse this information. One may worry that allowing discretion to managers in collecting and using information may have the perverse effect of increasing rather than reducing program leakage. Thus, a key dimension of heterogeneity surrounding the role of information will be the extent to which those who receive the information—be they supervisors or workers—have the incentives and ability to act on it.

### 5.2 Information flows and monitoring

Information on project and intermediaries' performance arises in multiple ways. The classic method remains via government auditing and inspection units that are required to monitor government programs. More recently, the rise of e-governance has meant that government agencies have access to large administrative data sets on funds flow, intermediaries' behavior (typically attendance), and monitored program outcomes. These data directly allow managers to obtain better real-time data on program performance and, in many cases, public availability of these data (aided in part by the rise of freedom of information acts) increases citizen monitoring.

In the following section, we first discuss findings from experiments that evaluate government monitoring processes and then we turn to citizen monitoring.

### 5.3 Government monitoring

#### 5.3.1 Does more information on performance improve outcomes?

Studies of audits. Several studies examine the role of government audits. [Olken \(2007\)](#) conducted a field experiment in Indonesia where a local village body implemented a

road construction program, and audits were conducted by the government agency. The key finding is that audits have a significant deterrence impact. Before villages began building road projects, some were randomly selected for a high-audit intensity group, where they faced an audit by the government agency with 100-percent probability as opposed to a 4-percent probability in the control group. [Olken \(2007\)](#) found substantial effects of the government audits, reducing unaccounted for expenditures by about 8-percentage points or about 30 percent from the baseline level.

[Ferraz and Finan \(2008\)](#) examined audits of municipal accounts in Brazil where small municipalities were randomly chosen to be audited by government auditors. They examine the impact of the timing of auditing on the probability that the mayor is reelected. They find that, conditioning on the actual number of corruption violations found by the auditors, those audited before the election were less likely to be reelected than those who were audited after the election for those with an intermediate number of violations.

An open question is whether a higher likelihood of punishment also had a deterrence effect in this setting. [Bobonis et al. \(2015\)](#) examine this question in the context of Puerto Rico. Puerto Rico has established an independent body that systematically conducts municipal government audits, the findings of which are made publicly available and disseminated to media sources. [Bobonis et al. \(2015\)](#) exploit two features of the audit process. First, municipalities are audited in a preestablished order, making the timing of audits and their dissemination predetermined. Second, audits are “timely audits,” such that reports released in the period leading up to an election are more likely to inform on the incumbent mayor’s activities than those reports published shortly after an election due to a high independent turnover rate of politicians. They find that timely audits induce a significant short-term reduction in municipal corruption levels of approximately 67 percent, as well as an increase in incumbent mayors’ electoral accountability. However, in contrast to these desirable short-run consequences of the audits, municipal corruption levels in the subsequent round of audits are, on average, the same in municipalities audited preceding the previous election and those whose audits became publicly available afterward. They also find that incumbent reelection rates in the subsequent election are significantly higher in municipalities in which there was an earlier timely audit. The presence of selection effects in future reelection rates, but not in corruption, is *prima facie* evidence in favor of the view that the information contained in the audits helps voters select competent but opportunistic politicians, rather than honest or virtuous ones.

There is, however, the potential for monitoring to backfire. [Lichand et al. \(2015\)](#) studies the introduction of the municipal audits as in [Ferraz and Finan \(2008\)](#) in a differences-in-differences framework. They show that in municipalities that expected they might be audited, procurement went down, with negative consequences for health. To the extent that the problems we observe are due to incompetence or laziness rather than corruption, as suggested by [Bandiera et al. \(2009\)](#), too much of a focus on corruption could backfire. We regard continued explorations of this issue as an important area for future work.

Biometric/time stamp studies. As e-governance becomes more widespread, countries have increased their investment in, and use of, e-monitoring systems. Several studies described in our discussion of incentives based on provider attendance ([Section 4](#)) utilize such systems. [Duflo et al. \(2012\)](#) based incentives on information obtained via time-stamped photographs. However, as the study by [Banerjee et al. \(2008\)](#) highlights, the robustness of such monitoring systems is sensitive to how tamper proof the monitoring mechanism is. If nurses can destroy the monitoring system (here, by literally breaking the time stamping mechanism), then they will do so. However, more importantly, if the incentive system allows for loopholes, then improved information may do little—a general lesson that exists above and beyond the nature of monitoring mechanism.

Two recent studies that expand our understanding of the issues at stake are [Dhaliwal and Hanna \(2014\)](#) and [Callen et al. \(2015\)](#). [Dhaliwal and Hanna \(2014\)](#) studied the rollout of biometric monitoring of the staff at primary health centers in South India. Health worker attendance increased by 14.7 percent in clinics with improved monitoring and was driven by lower-level staff in these centers (nurses and pharmacists), rather than by doctors. Improved monitoring had important health impacts: there was a 16-percent increase in the delivery of infants by a doctor and a 26-percent reduction in the likelihood that infants are born under 2500 grams. However, they also find lower staff satisfaction and widespread attempts by the staff to circumvent the system. Taken together, the results show both how improved monitoring can improve service delivery and also the rents at stake in the system. To the extent that staff dissatisfaction and delays in implementation tend to be more visible, this study also points to the importance of measuring impacts. In the absence of careful measurement of health outcomes, it would have been easy to focus on the partial implementation and to deem the system a failure.

If poor monitoring opens the door for shirking and potentially outright corruption by service providers, then we may expect these impacts to vary by the personality of the service provider. As discussed previously, [Callen et al. \(2015\)](#) report on a monitoring experiment where the traditional paper-based monitoring system for clinic utilization, resource availability, and worker absence was replaced by an Android smartphone application. In the new system, data generated by health inspections are transmitted to a central database using a mobile data connection in real time. Data are then aggregated and summary statistics, charts, and graphs are presented in a format designed in collaboration with senior health officials to effectively communicate information about health facility performance. Especially relevant given the study by [Dhaliwal and Hanna \(2014\)](#), these authors find that senior health inspectors who score higher on the Big Five personality inventory are more likely to respond to a report of an underperforming facility by compelling better subsequent staff attendance. More surprisingly, they also find that inspectors who score higher on personality tests are more likely to reduce absenteeism when dashboards are implemented. This paper provides one way into understanding the observed heterogeneity in responsiveness to better monitoring: individual personality characteristics.

### 5.3.2 Who collects information and does that matter?

A common concern in the literature is the veracity of information collected by monitors. For instance, both [Banerjee et al. \(2008\)](#) and [Dhaliwal and Hanna \(2014\)](#) report on how service providers seek to reduce the functionality of monitoring devices.

Incentives to provide poor-quality information may also arise if inspectors and auditors are corruptible by those they are intended to monitor. This possibility is, arguably, particularly stark in the case of private-sector auditors paid by the firms or institutions they audit. [Duflo et al. \(2013\)](#) examine the implications of corrupted information flows for regulatory efficacy. In a large field experiment conducted with the environmental regulator in Gujarat, they altered the assignment and payment mechanism for third-party environmental auditors of industrial plants. Under the status quo, the auditors were hired and paid for by the plant they audited. In the treatment group, auditors were instead randomly assigned to plants and paid a fixed salary from a central pool of funds. The experiment demonstrated that the status quo system was largely corrupted, with auditors systematically reporting plant emissions just below the standard, although true emissions were typically higher. Second, the treatment caused auditors to report more truthfully and significantly lowered the fraction of plants that were falsely reported as compliant with pollution standards. Third, treatment plants, in turn, reduced their pollution emissions.

A different margin of potential corruption in information acquisition is providing higher-level officials discretion in whom to inspect or audit and when. Another is the choice of intermediaries who have monitoring responsibilities.

[Duflo et al. \(2015\)](#) examine this issue in the context of environmental inspections in India. They examine whether raising the frequency of inspections in a rule-bound manner changes regulator behavior and improves plant compliance. They find that more inspections lead the regulator to send more warnings but not to increase incidence of punishments. They use detailed information on regulator plant interactions to show that the regulator uses his discretion to target information collection and punishment efforts at a smaller subset of highly polluting plants. In this case, regulatory discretion is valuable as it allows the regulator to best target his scarce inspection resources.

Finally, a different way to improving monitoring is by increasing private-sector incentives to report the truth. A commonly cited example is value added tax (VAT), which generates paper trails on transactions between firms. [Pomeranz \(2015\)](#) analyzes the role of third-party information for VAT enforcement through two randomized experiments among over 400,000 Chilean firms. She shows that announcing additional monitoring has less impact on transactions that are subject to a paper trail, indicating the paper trail's preventive deterrence effect. This leads to enforcement spillovers up the VAT chain. We return to the theme of how technology of service provision or revenue collection can be harnessed for better monitoring when we discuss the promise of e-governance.

## 5.4 Information flows and monitoring by citizens

The last decade has seen increased interest in monitoring undertaken directly by citizens. In part, this reflects the increasing incidence of freedom of information acts and in part the greater ease of obtaining already digitized data on program performance.

### 5.4.1 Does information on program performance matter?

[Björkman and Svensson \(2010\)](#) found that informing Ugandan citizens of the dismal state of local health service delivery and holding meetings between citizens and health workers to agree on action plans significantly reduced provider absenteeism, increased utilization, and improved health. In a second randomized evaluation, [Björkman et al. \(2014\)](#) examined a less expensive version of the program where they did not provide information on health worker performance and found no impacts, suggesting that access to information was key.

More recently, [Banerjee et al. \(2015\)](#) examine the impact of mailing cards with program information to beneficiaries of a subsidized rice program in Indonesia. They found that this increased the extent of subsidy receipt from the program. Beneficiaries received 26 percent more subsidy in treated villages. Ineligible households received no less, suggesting reduced leakage. The impact appears to be driven by citizens bargaining with local officials. Experimentally, adding the official price to the cards increased the subsidy by 21 percent compared to cards without price information. Additional public information increased higher-order knowledge about eligibility, leading to a 16-percent increase in subsidy compared to just distributing cards.

In contrast to these papers, [Banerjee et al. \(2010\)](#) report limited results from a report card intervention in which village volunteers prepared a report card on student learning. They interpret this as reflecting the absence of mechanisms to hold teachers accountable. This study points to the importance of understanding how monitoring mechanisms interact with underlying incentives for government workers. These ideas are explored in the study by [Pradhan et al. \(2014\)](#) who examine the role of school committees in improving education quality in Indonesia. The study is a randomized evaluation covering 520 schools in Central Java from 2007 to 2008. They have four main treatments. The first treatment facilitated democratic elections of school committee members. The second treatment linked school committees to the village council by facilitating joint planning meetings (which they describe as linkage). They benchmark these two treatments against more common treatments: providing block grants and providing training. Two years later, test scores increased by 0.17 standard deviations for linkage and 0.23 standard deviations for linkage plus elections. In contrast, training did not impact learning, and the effect of grants, while positive, was typically statistically indistinguishable from zero. Taken together, the contrasting results from India and Indonesia point strongly to the

importance of institutional reforms that lead to positive interactions between monitoring and incentive mechanisms.

### 5.5 Summary

The evidence reviewed in this section leads to several conclusions. First, monitoring can help resolve two agency problems. The first of these is the asymmetrical information that exists between the employer and the employee. With better monitoring, employees have less of an incentive to shirk, and employers also have the flexibility to offer high-powered contracts. The second is the information asymmetry between the service provider and the citizen. With better monitoring (and the disclosure of this information) citizens can hold their service providers accountable, by applying both bottom-up pressure, as well as inducing top-down pressure. Of course, improving monitoring capabilities is unlikely to induce much change without accountability mechanisms in place. Second, third-party reporting matters. With more monitoring, employees may try to game the system. Third-party reporting creates conflicting interests; this helps establish the veracity of the information.

## 6. TOWARDS SMART(ER) GOVERNANCE: THE PROMISE OF E-GOVERNANCE AND OTHER AVENUES

A common theme that emerges from the body of experimental evidence is the sensitivity of individual behavior to the incentives they face. This, in turn, points to the importance of the structure within which government workers function. Below, we discuss how an emerging body of evidence suggests that e-governance and other technological changes in how the government functions may well help improve governance in low-income settings, perhaps by replacing some of the functions played by personnel—who are subject to all the various problems discussed previously—with technological solutions.

There are a number of ways in which technology can constrain the discretion of local officials and improve performance. For example, in low-income countries the rural poor—an important target group for government transfer programs—are often less informed about state services available to them. In addition, limited state presence in rural areas implies that it is often harder to deploy traditional personnel-intensive monitoring mechanisms to ensure that intended program beneficiaries get their due. As a result, traditional modes of delivery tend to provide the village-level service provider significant discretion in who gets the transfer and when. In cases where payment is supposed to be conditional on the beneficiary undertaking certain activities (e.g., working in a workfare program or children going to school), the village provider often receives funds ahead of the activity having occurred. The monitoring system in these cases often focuses on reviewing the

funds request system. Two recent papers examine how e-governance and the use of biometrics can help improve both the fund-flow system from central coffers to village-level coffers and the transfer of resources from the village-level provider to the final beneficiaries.

[Banerjee et al. \(2015\)](#) report on a field experiment which evaluated an e-governance reform of the fund-flow system for the federal welfare program in the Indian state of Bihar. The reform changed the traditional fund-flow practice by instead conditioning fund disbursement for wage payments on incurred expenditure as reflected in worker detail entry on a new electronic platform. This reform reduced the number of administrative tiers associated with wage disbursement and changed the informational requirements for requesting and disbursing program funds. It did not alter the flow of funds from the village fund to workers. The authors find that program expenditure and reported employment declined by 25% but with no discernible impact on actual employment as measured by an independent survey. Thus, the financial reform was effective in reducing corruption and program costs, but actual demand that was met by the program was unchanged.

[Muralidharan et al. \(2014\)](#) provide evidence on the last step of the payment process. They evaluate the impact of a biometrically authenticated payments infrastructure which provided “Smartcards” to beneficiaries of the federal welfare program. Their experiment randomized the rollout of Smartcards over 158 subdistricts and 19 million people in the southern state of Andhra Pradesh. They find that, while incompletely implemented, the new system delivered a faster, more predictable, and less corrupt National Rural Employment Guarantee Scheme (NREGS) payments process without adversely affecting program access. For each of these outcomes, treatment group distributions first-order stochastically dominated those of the control group. The investment was cost-effective, as time savings to NREGS beneficiaries alone were equal to the cost of the intervention, and there was also a significant reduction in the “leakage” of funds between the government and beneficiaries in both NREGS and Social Security Pension (SSP) programs.

[Barnwal \(2014\)](#) reports supportive quasi-experimental evidence for the staggered rollout and subsequent pulling back of a biometric-based scheme for fuel subsidies in India. He finds that the biometric-based transfer policy reduced fuel purchases in the domestic fuel sector by 11–14% suggesting a reduction in subsidy diversion. In addition, after the policy is terminated, fuel purchases in the domestic sector revert to levels similar to before the policy was introduced.

A different way in which technology can help delivery of transfer programs is the use of mobile money. [Aker et al. \(2014\)](#) use data from a randomized experiment of a mobile money cash transfer program in Niger and find evidence of benefits of this new system: Households receiving mobile transfers had higher diet diversity and children consumed more meals per day. These results can be partially attributed to increased time saving, as

m-transfer program recipients spent less time traveling to and waiting for their transfer, as well as increased intrahousehold bargaining power for women.

Technology can also play a role in reducing malfeasance in elections and promoting electoral accountability. [Callen and Long \(2015\)](#) implement an experiment to estimate the causal effects of photo quick count—a technology used to reduce electoral fraud—and its announcement on aggregation fraud. Photo quick count announcement reduced damaging of election materials by candidate representatives from 18.9 to 8.1 percent, and reduced votes for politically powerful candidates at a given polling location from about 21 to about 15 percent.

In a quasi-experimental study, [Fujiwara \(2015\)](#) studies the introduction of electronic voting on voter enfranchisement in Brazil. Electronic voting was introduced at scale in 1998 elections. But because of a limited supply in devices, only municipalities with more than 40,500 registered voters used the new technology. Using a regression discontinuity design, he finds that electronic voting reduced residual voting in state legislature elections by a magnitude larger than 10% of total turnout, thus enfranchising millions of voters. By enfranchising a poorer and less education population, the introduction of electronic voting led to an increase in the number of prenatal visits by health professionals and lowered the prevalence of low-weight births (below 2500 g) by less educated women.

A third area where technology shows some promise is in government procurement. Government procurement accounts for an enormous amount of government expenditures, and despite many regulations put in place to ensure that procurement is conducted fairly and with limited corruption, the fact that procurement regulations must be implemented by officials allows scope for discretion. For example, officials can withhold detailed bidding documents from bidders outside the favored cartels. [Lewis-Faupel et al. \(2016\)](#) study the introduction of electronic procurement systems for public works projects in India and Indonesia using a differences-in-differences design that takes advantage of the differential rollout of electronic procurement by states/provinces over time. These electronic procurement systems allow greater access to information for all potential bidders and ensure that the procurement rules are followed correctly. [Lewis-Faupel et al. \(2016\)](#) find that electronic procurement leads to contracts being more likely to be won by providers from outside the region where the project is being executed, suggesting that an important role for e-procurement is increasing access to information. They also find that it leads to quality improvements, though not lower prices paid by the government.

## 7. CONCLUDING THOUGHTS

In countries where the quality of government is low, public servants tend to be paid relatively well. While it is difficult to assign causality to this relationship, it does hint at some important directions for future research.

That public servants earn on average significantly more than their private sector counterparts does not suggest that financial incentives do not matter for bureaucratic performance. The evidence we have reviewed here suggests that they clearly do. But rather than differences in the levels, it is the nature of the incentives that perhaps matter most for performance. In most countries, wages in the public sector tend to exhibit a high degree of compression, which combined with long tenure rates can make it hard to incentivize an individual once employed. But in settings where government can offer high-powered incentives (e.g., tax administration or education), the evidence suggests that public servants do in fact perform better.

High-powered incentives come with tradeoffs: these incentive schemes have the potential to discourage effort among the prosocially motivated, as well as to create issues of multitasking, though the empirical evidence suggests that in many cases these concerns are not as strong as one might initially have thought. The literature has begun to identify some of these tradeoffs, but much more research is needed to better understand in which settings these issues are most likely to arise. For example, task complexity might provide such a setting. The multitasking concerns associated with performance pay are more likely to arise when bureaucrats are tasked with complex jobs. At the same time, complex jobs are more difficult to monitor. Whether the benefits of lower monitoring costs outweigh the costs associated with multitasking is an interesting question with important implications for how bureaucracies should be organized.

Financial incentives also matter for selection. Organizations that offer higher wages will attract more qualified applicants. Given the large public-sector wage gaps, one might be tempted to reason that selection issues are not a first-order concern. But high wages are only a necessary condition for attracting talent, not a sufficient one. The type of individuals who work in the public sector will ultimately depend on how candidates are screened. The screening mechanism can easily undo any positive selection created by higher wages, as is perhaps the case when countries hire based on patronage and not for meritocratic reasons. Governments vary in the way they screen their public servants. To understand how governments should screen among the candidates who apply to further improve selection is another exciting avenue for future research. It is of course difficult to consider optimal screening mechanisms without raising the question of what personality traits we should screen for. While there is an extensive empirical literature arguing that individuals who exhibit high degrees of prosocialness perform better in the public sector, experimental evidence of this relationship is virtually nonexistent.

The usefulness of financial incentives can also be limited if internal accountability mechanisms do not exist or function. One issue that bureaucracies typically face is the inability to perfectly monitor their employees. Fortunately advances in technology appear to be a step forward. The use of smartphone technology and e-governance platforms not only promote transparency and accountability but also serve as disciplining devices. Importantly, these technological advances may also create a feedback loop on the

compensation structure of employees. As governments increasingly adopt these new technologies, thus enabling them to better monitor and evaluate its employees, the set of contracts that it can offer its employees expands. The relationship between technology adoption and compensation scheme is another exciting area of future research.

As this survey documents, the (experimental) research on trying to understand how bureaucracies work is still in its infancy, so there is plenty to do and a lot to learn. We are excited to see what the next several years will bring for this research agenda, which has the potential to unlock some of the doors to efficient service delivery and good governance.

## APPENDIX

Region	Country	Year	Survey	Source
Africa	Ghana	2013	STEP Skills Measurement Program survey	World Bank
Africa	Kenya	2013	STEP Skills Measurement Program survey	World Bank
Africa	Nigeria	2012	Living Standards Measurement survey	World Bank
Africa	Niger	2011	Living Standards Measurement survey	World Bank
Africa	Malawi	2010	Living Standards Measurement survey	World Bank
Africa	Tanzania	2010	Living Standards Measurement survey	World Bank
Africa	Timor-Leste	2007	Living Standards Measurement survey	World Bank
Asia	Lao PDR	2012	STEP Skills Measurement Program survey	World Bank
Asia	Sri Lanka	2012	STEP Skills Measurement Program survey	World Bank
Asia	Vietnam	2012	STEP Skills Measurement Program survey	World Bank
Asia	India	2011	Socio Economic survey	National Sample Survey Office, Government of India
Asia	Korea, Rep.	2011	Korean General Social Survey	ICPSR

*Continued*

<b>Region</b>	<b>Country</b>	<b>Year</b>	<b>Survey</b>	<b>Source</b>
Asia	Indonesia	2007	Indonesia Family Life Survey	RAND Corporation
Asia	Pakistan	2006	Labour Force Survey	Federal Bureau of Statistics, Government of Pakistan
Central Asia	Armenia	2013	STEP Skills measurement Program survey	World Bank
Central Asia	Georgia	2013	STEP Skills Measurement Program survey	World Bank
Central Asia	Tajikistan	2009	Living Standards Measurement survey	World Bank
Europe	United Kingdom	2014	Quarterly Labour Force Survey	UK data service
Europe	Albania	2011	Labour Force Survey	Institute of Statistics (INSTAT), Republic of Albania
Europe	Bulgaria	2007	Living Standards Measurement survey	World Bank
Europe	Serbia	2007	Living Standards Measurement survey	World Bank
Europe	Bosnia-Herzegovina	2004	Living Standards Measurement survey	World Bank
Latin America	Argentina	2014	Permanent Household Survey	National Institute of Statistics and Census, Republic Argentina
Latin America	Mexico	2014	National Survey of Occupation and Employment	National Institute of Statistics and Geography (INEGI), Government of Mexico
Latin America	Bolivia	2012	STEP Skills Measurement Program survey	World Bank
Latin America	Colombia	2012	STEP Skills Measurement Program survey	World Bank
Latin America	Peru	2011	Specialized household survey on employment levels	Ministry of Work and Employment Promotion, Government of Peru

Region	Country	Year	Survey	Source
Latin America	Panama	2008	Living Standards Measurement survey	World Bank
Latin America	Nicaragua	2005	Living Standards Measurement survey	World Bank
Middle East	Egypt	2012	Egypt Labor Market Panel Survey	Economic Research Forum
Middle East	Iraq	2006	Living Standards Measurement survey	World Bank
North America	United States	2010	Current Population Survey	National Bureau of Economic Research (NBER)

## REFERENCES

- Aker, J., Boumnijel, R., McClelland, A., Tierney, N., 2014. Payment Mechanisms and Anti-poverty Programs: Evidence from a Mobile Money Cash Transfer Experiment in Niger. Working Paper 268.
- Almlund, M., Lee Duckworth, A., Heckman, J., Kautz, T., 2011. Personality psychology and economics. In: Hanushek, E.A., Machin, S., Woessmann, L. (Eds.), *Handbook of the Economics of Education*. Elsevier, Amsterdam, pp. 1–181.
- Aronson, E., 1999. The power of self-persuasion. *Am. Psychol.* 54 (11), 875–884.
- Ashraf, N., Bandiera, O., Kelsey Jack, B., 2014. No margin, No mission? a field experiment on incentives for public service delivery. *J. Public Econ.* 120, 1–17.
- Ashraf, N., Bandiera, O., Lee, S.S., 2015. Do-gooders and Go-getters: Career Incentives, Selection, and Performance in Public Service Delivery (Working Paper).
- Baicker, K., Jacobson, M., 2007. Finders keepers: forfeiture laws, policing incentives, and local budgets. *J. Public Econ.* 91, 2113–2136.
- Bandiera, O., Prat, A., Valletti, T., 2009. Active and passive waste in government spending: evidence from a policy experiment. *Am. Econ. Rev.* 99 (4), 1278–1308.
- Banerjee, A.V., Banerji, R., Duflo, E., Glennerster, R., Khemani, S., 2010. Pitfalls of participatory programs: evidence from a randomized evaluation in education in India. *Am. Econ. J. Econ. Policy* 2 (1), 1–30.
- Banerjee, A.V., Chattopadhyay, R., Duflo, E., Keniston, D., Singh, N., 2014. Can Institutions Be Reformed from within? Evidence from a Randomized Experiment with the Rajasthan Police. NBER Working Paper 17912.
- Banerjee, A.V., Hanna, R., Kyle, J., Olken, B., Sumarto, S., 2015. The Power of Transparency: Information, Identification Cards and Food Subsidy Programs in Indonesia (HKS Faculty Research Working Paper Series).
- Banerjee, A.V., Glennerster, R., Duflo, E., 2008. Putting a band-aid on a corpse: incentives for nurses in the Indian public health care system. *J. Eur. Econ. Assoc.* 6 (2/3), 487–500.
- Banerjee, R., Baul, T., Rosenblat, T., 2015. On self selection of the corrupt into the public sector. *Econ. Lett.* 127, 43–46.
- Banuri, S., Keefer, P., 2013. Intrinsic Motivation, Effort and the Call to Public Service. World Bank Policy Research Working Paper 6729.
- Barnwal, P., 2014. Curbing Leakage in Public Programs with Biometric Identification Systems: Evidence from India's Fuel Subsidies (Job Market Paper).
- Basinga, P., Gertler, P., Binagwaho, A., Soucat, A., Sturdy, J., Vermeersch, C., 2011. Effect on maternal and child health services in rwanda of payment to primary health-care providers for performance: an impact evaluation. *Lancet* 377 (9975), 1421–1428.

- Becker, G.S., Stigler, G.J., 1974. Law enforcement, malfeasance, and compensation of enforcers. *J. Leg. Stud.* 1–18.
- Bellé, N., 2012. Experimental evidence on the relationship between public service motivation and job performance. *Public Adm. Rev.* 73 (1), 143–153.
- Bénabou, R., Tirole, J., 2006. Incentives and prosocial behavior. *Am. Econ. Rev.* 96 (5), 1652–1678.
- Besley, T., Ghatak, M., 2005. Competition and incentives with motivated agents. *Am. Econ. Rev.* 95 (3), 616–636.
- Björkman, M., de Walque, D., Svensson, J., 2014. Information Is Power: Experimental Evidence on the Long-Run Impact of Community Based Monitoring. World Bank Policy Research Working Paper 7015.
- Björkman, M., Svensson, J., 2010. When is community-based monitoring effective? Evidence from a randomized experiment in primary health in Uganda. *J. Eur. Econ. Assoc.* 8 (2/3), 571–581.
- Bloom, N., Van Reenen, J., 2007. Measuring and explaining management practices across firms and countries. *Q. J. Econ.* 122 (4), 1351–1408.
- Bloom, N., Eifert, B., Mahajan, A., McKenzie, D., Roberts, J., 2013. Does management matter? Evidence from India. *Q. J. Econ.* 128 (1), 1–51.
- Bobonis, G.J., Cámara Fuertes, L.R., Schwabe, R., 2015. Monitoring Corruptible Politicians (Working Paper).
- Borjas, G.J., 2015. The Wage Structure and the Sorting of Workers into the Public Sector. NBER Working Paper 9313.
- Boycko, M., Shleifer, A., Vishny, R., 1996. A theory of privatisation. *Econ. J.* 106, 309–319.
- Burgess, S., Propper, C., Ratto, M., von Scholder, S.H.K., Tominey, E., 2010. Smarter task assignment or greater effort: the impact of incentives on team performance. *Econ. J.* 120, 968–989.
- Callen, M., Gulzar, S., Hasanain, A., Khan, Y., Rezaee, A., 2015. Personalities and Public Sector Performance: Evidence from a Health Experiment in Pakistan. NBER Working Paper 21180.
- Callen, M., Long, J., 2015. Institutional corruption and election fraud: evidence from a field experiment in Afghanistan. *Am. Econ. Rev.* 105 (1), 354–381.
- Chaudhury, N., Hammer, J., Kremer, M., Muralidharan, K., Halsey Rogers, F., 2006. Missing in action: teacher and health worker absence in developing countries. *J. Econ. Perspect.* 20 (1), 91–116.
- Church, T.W., Heumann, M., 1989. The underexamined assumptions of the invisible hand: monetary incentives as policy instruments. *J. Policy Anal. Manag.* 8 (4), 641–657.
- Cowley, E., Smith, S., 2014. Motivation and mission in the public sector: evidence from the world values survey. *Theory Decis.* 76, 241–263.
- Dal Bó, E., Finan, F., Rossi, M.A., 2013. Strengthening state capabilities: the role of financial incentives in the call to public service. *Q. J. Econ.* 128 (3), 1169–1218.
- Delfgaauw, J., Dur, R., 2007. Signaling and screening of workers' motivation. *J. Econ. Behav. Organ.* 62 (4), 605–624.
- Deserranno, E., 2015. Financial Incentives as Signals: Experimental Evidence from the Recruitment of Health Workers (Working Paper).
- Dhaliwal, I., Hanna, R., 2014. Deal with the Devil: The Successes and Limitations of Bureaucratic Reform in India. NBER Working Paper 20482.
- Dizon-Ross, R., Dupas, P., Robinson, J., 2015. Governance and the Effectiveness of Public Health Subsidies. SCID Working Paper 510.
- Dohmen, T., Falk, A., 2010. You get what you pay for: incentives and selection in the education system. *Econ. J.* 120 (546), F256–F271.
- Duflo, E., Greenstone, M., Pande, R., Ryan, N., 2013. Truth telling by third-party audits and the response of polluting firms: experimental evidence from India. *Q. J. Econ.* 128 (4), 1499–1545.
- Duflo, E., Greenstone, M., Pande, R., Ryan, N., 2015. The Value of Regulatory Discretion: Estimates from Environmental Inspections in India. Harvard Environmental Economics Program Discussion Paper 15–60.
- Duflo, E., Hanna, R., Ryan, S.P., 2012. Incentives work: getting teachers to come to school. *Am. Econ. Rev.* 102 (4), 1241–1278.

- Evans, P.B., 1995. Embedded autonomy: states and industrial transformation. Vol. 25. Princeton University Press, Princeton, NJ.
- Evans, P., Rauch, J.E., 1999. Bureaucracy and growth: a cross-national analysis of the effects of "Weberian" state structures on economic growth. *Am. Sociol. Rev.* 64 (5), 748–765.
- Ferraz, C., Finan, F., 2008. Exposing corrupt politicians: the effects of Brazil's publicly released audits on electoral outcomes. *Q. J. Econ.* 123 (2), 703–745.
- Francois, P., 2000. Public service motivation' as an argument for government provision. *J. Public Econ.* 78 (3), 275–299.
- Fujiwara, T., 2015. Voting technology, political responsiveness, and infant health: evidence from Brazil. *Econometrica* 83 (2), 423–464.
- Gertler, P., Vermeersch, C., 2012. Using Performance Incentives to Improve Health Outcomes. World Bank Policy Research Working Paper 6100.
- Glewwe, P., Illias, N., Kremer, M., 2010. Teacher incentives. *Am. Econ. J. Appl. Econ.* 2 (3), 205–227.
- Gneezy, U., Rustichini, A., 2000. A fine is a price. *J. Leg. Stud.* 29 (1).
- Grant, A.M., Campbell, E., Chen, G., Cottone, K., Lapedis, D., Lee, K., 2007. Impact and the art of motivation maintenance: the effects of contact with beneficiaries on persistence behavior. *Organ. Behav. Hum. Decis. Process.* 103, 53–67.
- Hanna, R., Wang, S.-Y., 2014. Dishonesty and Selection into Public Service: Evidence from India. NBER Working paper 19649.
- Holmstrom, B., Milgrom, P., 1987. Aggregation and linearity in the provision of intertemporal incentives. *Econometrica* 55 (2), 303–328.
- Iyer, L., Mani, A., 2012. Traveling agents: political change and bureaucratic turnover in India. *Rev. Econ. Statistics* 94 (3), 723–739.
- Khan, A.Q., Khwaja, A.I., Olken, B.A., 2014. Tax Farming Redux: Experimental Evidence on Performance Pay for Tax Collectors. NBER Working Paper 20627.
- Kahn, C.M., Silva, E.C.D., Ziliak, J.P., 2001. Performance-based wages in tax collection: the Brazilian tax collection reform and its effects. *Econ. J.* 111 (468), 188–205.
- Lagarde, M., Blaauw, D., 2014. Pro-social preferences and self-selection into rural jobs: evidence from South African nurses. *J. Econ. Behav. Organ.* 107 (Part A), 136–152.
- Lazear, E.P., Oyer, P., 2012. Personnel economics. In: Gibbons, R., Roberts, J. (Eds.), *The Handbook of Organizational Economics*. Princeton University Press, pp. 479–519.
- Lichand, G., Lopes, M.F.M., Medeiros, M., 2015. Is Corruption Good for Your Health? (Working Paper).
- Lewis-Faupel, S., Neggers, Y., Olken, B., Pande, R., 2016. Can Electronic Procurement Improve Infrastructure Provision? Evidence from Public Works in India and Indonesia. *AEJ: Economic Policy* 8 (3), 258–283.
- Manning, A., 2011. Imperfect competition in the labor market. In: Ashenfelter, O., Card, D. (Eds.), *Handb. Labor Econ.* 4, 973–1041.
- Meares, T.L., 1995. Rewards for Good Behavior: Influencing Prosecutorial Discretion and Conduct with Financial Incentives. Faculty Scholarship Series Paper 473.
- Mellstrom, C., Johannesson, M., 2008. Crowding out in blood donation: was titmuss right? *J. Eur. Econ. Assoc.* 6 (4), 845–863.
- Miller, G., Luo, R., Zhang, L., Sylvia, S., Shi, Y., Foo, P., Zhao, Q., Martorell, R., Medina, A., Rozelle, S., 2012. Effectiveness of provider incentives for anaemia reduction in rural China: a cluster randomised trial. *BMJ* 345, e4809.
- Miller, J.M., Selva, L.H., 1994. Drug enforcement's double-edged sword: an assessment of asset forfeiture programs. *Justice Q.* 11 (2), 313–335.
- Muralidharan, K., 2012. Long-term Effects of Teacher Performance Pay: Experimental Evidence from India (Working Paper).
- Muralidharan, K., Sundararaman, V., 2011. Teacher performance pay: experimental evidence from India. *J. Political Econ.* 119 (1), 39–77.
- Muralidharan, K., Niehaus, P., Sukhtankar, S., 2014. Building State Capacity: Evidence from Biometric Smartcards in India. NBER Working Paper 19999.

- Naff, K.C., Crum, J., 1999. Working for America: does public service motivation make a difference? *Rev. Public Personnel Adm.* 19 (4), 5–16.
- Neal, D., 2011. The design of performance pay in education. In: Hanushek, E., Machin, S., Woessmann, L. (Eds.), *Handbook of Economics of Education*, vol. 4. Elsevier, Oxford, pp. 495–550.
- Olken, B.A., 2007. Monitoring corruption: evidence from a field experiment in Indonesia. *J. Public Econ.* 115 (2), 200–249.
- Olken, B.A., Onishi, J., Wong, S., 2014. Should aid reward performance? evidence from a field experiment on health and education in Indonesia. *Am. Econ. J. Appl. Econ.* 6 (4), 1–34.
- Park, S.M., Rainey, H.G., 2008. Leadership and public service motivation in U.S. Federal agencies. *Int. Public Manag. J.* 11 (1), 109–142.
- Parrillo, N., 2013. *Against the Profit Motive: The Salary Revolution in American Government, 1780–1940*. Yale University Press.
- Perry, J.L., 1996. Measuring public service motivation: an assessment of construct reliability and validity. *J. Public Adm. Res. Theory* 6 (1), 5–22.
- Perry, J.L., Hondeghem, A., 2008. Building theory and empirical evidence about public service motivation. *Int. Public Manag. J.* 11 (1), 3–12.
- Petrovsky, N., 2009. Does public service motivation predict higher public service performance? A research synthesis. *J. Public Adm. Res. Theory* 20 (Suppl. 2), i261–i279.
- Pomeranz, D., 2015. No taxation without information: deterrence and self-enforcement in the value added tax. *Am. Econ. Rev.* 105 (8), 2539–2569.
- Pradhan, M.P., Suryadarma, D., Beatty, A., Wong, M., Gaduh, A., Alisjahbana, A., Artha, R.P., 2014. Improving educational quality through enhancing community participation: results from a randomized field experiment in Indonesia. *Am. Econ. J. Appl. Econ.* 6 (2), 105–126.
- Prendergast, C., 2007. The motivation and bias of bureaucrats. *Am. Econ. Rev.* 97 (1), 180–196.
- Rasul, I., Rogger, D., 2015. Management of Bureaucrats and Public Service Delivery: Evidence from the Nigerian Civil Service (Working Paper).
- Ritz, A., 2009. Public service motivation and organizational performance in Swiss federal government. *Int. Rev. Adm. Sci.* 75 (1), 53–78.
- Shleifer, A., Vishny, R.W., 1994. Politicians and firms. *Q. J. Econ.* 109 (4), 995–1025.
- Steijn, B., 2008. Person-environment fit and public service motivation. *Int. Public Manag. J.* 11 (1), 13–27.

## CHAPTER 7

# Designing Social Protection Programs: Using Theory and Experimentation to Understand How to Help Combat Poverty

R. Hanna\*,<sup>1</sup> D. Karlan<sup>§</sup>

\*Harvard University, Cambridge, MA, United States

§Yale University, New Haven, CT, United States

<sup>1</sup>Corresponding author: E-mail: Rema\_Hanna@hks.harvard.edu

## Contents

1. Introduction	516
2. Redistributive Programs	519
2.1 How to target the poor?	520
2.1.1 Targeting methods	521
2.1.2 Experimentally testing between targeting methods	523
2.2 Evaluating the impacts of redistributive programs	526
3. Missing Insurance Markets	528
4. Behavioral Constraints	530
4.1 Providing in-kind transfers to constrain spending choices	530
4.1.1 The rationale for in-kind programs	530
4.1.2 Evaluating in-kind programs	532
4.2 Adding conditions to incentivize behavior	534
4.2.1 Evaluating conditions relative to basic redistributive programs	534
4.2.2 Which conditions should be imposed?	536
4.2.3 Enforcing the conditions	536
5. Market Failures Preventing Asset Accumulation	537
5.1 Building productive assets	538
5.2 Building long-term financial assets—pensions	542
6. Ideas Only Go So Far: Implementation Matters Too	542
7. Conclusion: Key Areas for Further Work	545
7.1 Key areas for further work	546
7.1.1 Interactions of demand and supply	546
7.1.2 Long-term effects	546
7.1.3 General equilibrium effects	547
7.2 Final thoughts	548
References	548

## Abstract

"Antipoverty" programs come in many varieties, ranging from multifaceted, complex programs to more simple cash transfers. Articulating and understanding the root problem motivating government and nongovernmental organization intervention are critical for choosing among many antipoverty policies or combinations thereof. Policies should differ depending on whether the underlying problem is about uninsured shocks, liquidity constraints, information failures, or some combination of all of the above. Experimental designs and thoughtful data collection can help diagnose the root problems better, thus providing better predictions for what antipoverty programs to employ in specific conditions and contexts. However, the more complex theories are likewise more challenging to test, requiring larger samples, and often more nuanced experimental designs, as well as detailed data on many aspects of household and community behavior and outcomes. We provide guidance on these design and testing issues for social protection programs, from how to target programs, to who should implement the program, and to whether and what conditions to require for program participation. In short, careful experimentation-designed testing can help provide a stronger conceptual understanding of why programs do or not work, thereby allowing one to ultimately make stronger policy prescriptions that further the goal of poverty reduction.

## Keywords

Social protection; Development; Antipoverty

## JEL Codes

O10; O12; H53

## 1. INTRODUCTION

In low-income countries, more than one billion individuals are enrolled in at least one safety net program (Gentilini et al., 2014).<sup>1</sup> These programs come in various forms and sizes. Some aim to simply supplement consumption in hard times. Other newer, more nuanced, social protection programs aim to address the underlying market failures that may have contributed to a household's persistent state of poverty in the first place, driven by a belief that directly addressing these failures may help families break out of a poverty trap. The ultimate choice of program—or combination therein—that countries choose to implement will depend greatly on their social goals, institutional capabilities, and resources. However, even within each broad category of program, the specific design choices made and methods of implementation may affect whether these programs actually achieve their stated goals.

To start thinking about how to design—and then test the impacts of—safety net programs, we begin by classifying them into four main categories based on the underlying motivation for intervening. The first and simplest category is comprised of programs designed for redistributive purposes, e.g., recognizing that the marginal utility of

<sup>1</sup> Throughout this chapter we refer to "social protection" and "safety net" programs as one and the same.

consumption is higher for the poor than for the rich, and thus transfers are socially optimal from a utilitarian perspective (assuming naturally that the taxation process does not create considerable deadweight losses). While these programs can differ in actual design, they share the common feature of first identifying the neediest families along a particular metric and then providing them with cash transfers. Of course, these programs could still have long-run growth impacts, e.g., if they are large enough to provide sufficient capital to start new agricultural, migratory, or business activities (e.g., [Banerjee and Newman, 1993](#); [McKenzie and Woodruff, 2006](#)) or if they persuade risk averse households to invest in riskier but more profitable endeavors ([Chetty and Looney, 2006](#)) and so forth.<sup>2</sup> But, the primary goal of these types of programs is simply to limit poverty and hunger by ensuring that households attain a minimum living standard.

Second, a missing insurance market may motivate a social protection program. The poor faces many risks, such as unexpected health costs, agricultural damage, or job losses. Without fully functioning financial markets, households may be unable to borrow to smooth consumption. Informal credit and insurance markets provide another avenue to do so, but they often underperform and have become even less effective as countries grow and urbanization breaks down traditional social networks ([Coady, 2004](#)). Furthermore, even if households smooth consumption in the face of income shocks, they may be doing so in the short run by making long-term sacrifices, such as pulling children out of school. The insurance market may exist but needs a subsidy to increase quantity demanded, or it may not exist at all and the safety net program directly provides protection in bad times. Social protection from natural disasters is an extreme example of the insurance motivation; however, economic theory and empirical work have had less to say about the structure of such policies. Finally, unemployment and health insurance programs also fit into this often-(at least partially) missing insurance market category.

Third, there may be behavioral or household-bargaining constraints that influence the choices of low-income households and perpetuate poverty. For example, difficulty in resisting immediate temptations can lead to undersaving and thus underinvestment in lumpy goods. Intrahousehold-bargaining issues can lead to suboptimal outcomes for the underpowered, typically women and thus children. Many transfer programs aim to account for these behavioral factors, by providing in-kind transfers (such as food) rather than cash to prevent temptation or by providing transfers to women rather than men to increase female-bargaining powering in the household. Importantly, many programs—conditional cash or in-kind programs—even directly condition assistance to poor families

<sup>2</sup> Notable examples that have found changes increase in investment as a result of cash transfer programs include [Covarrubias et al. \(2012\)](#) who find that Malawi's Social Cash Transfer led to an increase in agricultural investments; and [Gertler et al. \(2012\)](#) who find that Mexico's CCT (Progresa) led to higher levels of agricultural income, as households partially invested a portion of the cash in productive assets.

on behaviors that society would like to address, e.g., children attending school and receiving vaccinations or other preventive health measures. Finally, “workfare” programs also share this aim, by providing transfers conditional on labor force training and/or participation.

Fourth, and finally, there may be market failures preventing asset accumulation. We focus on two domains where asset accumulation could be important for social protection. First, short- and medium-run productive asset building could increase household income, thus allowing individuals to no longer need consumption transfers from government. Rather than cash transfers, productive asset transfers as well as training, coaching, and informational programs may be critical to motivate and improve investment choices to make such a goal attainable. Second, long-term financial asset building may be difficult if savings markets are missing, or if behavioral and household constraints discussed earlier bind. The current savings market infrastructure does not facilitate pension savings for individuals, and building such markets could be critical for improving consumption for the elderly in developing countries, just as in developed countries.

Naturally many programs can and often do target multiple issues at once, for two reasons. First, individuals may face multiple market failures, thus motivating a more complex program. For example, recently multifaceted programs have arisen that explicitly aim to “graduate” households out of extreme poverty by providing households with different inputs, including working capital, assets, and jobs training. These programs aim to increase household earnings capacity by concurrently relieving a number of different barriers to economic growth. Second, low-income households even within a particular setting may face different market failures and thus the optimal policy may not be the same for all. This difference across households may motivate targeting different aspects of a program to different people or designing programs that manage to address different issues for different people.

In short, a wide range of tools are available to policy-makers in the goal of poverty alleviation. This variety naturally causes us to question: what are the right programs, who should they be targeted to, and how do we know that they are working? Randomized evaluations can answer this question by providing clear answers to whether or not a program “works.” However, importantly, a randomized evaluation can go even further, providing insights into why it works, i.e., the underlying mechanisms that drive the observed treatment effects. By doing so, the evaluation can offer greater insight into whether a similar program would work elsewhere or how the program should change as circumstances change even within the same context. To generate these insights, a well-designed evaluation must pay careful attention to the theory behind the different forms of social protection programs, consider multiple treatment variations to isolate theoretical channels, and be creative in data collection.

We will first discuss issues to consider in testing programs that fall under each of the four categories we lay about earlier, weaving in examples and knowledge from the current state

of the literature.<sup>3</sup> We then address issues pertaining to implementation, as success not only relies on whether the proposed program theoretically can achieve the social goal it was designed to address but also relies on how it is implemented in practice. Finally, we offer advice on further research needed in this space, including a better understanding of both the general equilibrium and long-run impacts of these programs.

## 2. REDISTRIBUTIVE PROGRAMS

Motivations for redistribution, absent specific market failures, inevitably come from social preferences for lower levels of inequality or from philosophical tenets such as utilitarianism. In the simplest utilitarian form, elegantly put forward by Peter Singer (1997), redistribution is motivated by an awareness of the stark trade-offs between more of one's own consumption (for the wealthy) versus more consumption for the poor. Philosophical motivations abound, naturally: for example, John Rawls' *A Theory of Justice* (1971) argues for a veil of ignorance in which one makes moral and policy judgments without knowledge of one's own position in society. This hypothetical construct leads to arguments for redistribution to the worst-off members of society.

With even minimal weighting on other people's utility in one's own utility, some redistribution typically becomes the individually optimal policy for all. This naturally motivates why an individual may redistribute their wealth to others (i.e., through charitable actions), but not necessarily why a government, through taxation, may effectively mandate such redistribution. However, there are many reasons why a government may mandate such redistribution, rather than leave it to the voluntary actions of individuals. First, transaction costs may be high for individuals to target the poorest effectively. Second, it may be costly for wealthy individuals to transfer wealth to the poorest. Third, high levels of inequality can cause social strife, or even violent uprising, and collective action problems make it difficult for a purely voluntary system to generate sufficient redistribution to achieve the social optimal. Fourth, behavioral theories could explain why individuals underredistribute if left to their own device but do have a stated preference for more redistribution. This is akin to models of quasi-hyperbolic preferences or the dual self (Fudenberg and Levine, 2006) applied to redistribution: if one is maximizing the welfare of the more deliberative and sharing self (as opposed to the impulsive and selfish self), then one may want a commitment device to help stay in line with their more deliberative preferences. A government program for redistribution thus becomes such a device. Fifth,

<sup>3</sup> Summarizing the broad and sizable literature on poverty alleviation programs poses a unique set of challenges. We have tried to focus on key ideas and use the literature to help provide examples when possible, as well as to give insight into where open questions persist. We have tried to cover the important papers in the literature but cannot cover all of them while keeping this chapter a reasonable length. We apologize in advance to those whose papers that we do not cover in detail.

individuals may be misinformed about the current level of inequality and poverty, and their relative wealth, whereas policy-makers may be more informed. In the United States, for example, recent work shows that most people vastly underestimate the level of inequality in the United States and state a preference for more equality, even though they are simultaneously opposed to more taxes. This finding demonstrates a clear information gap (Norton and Ariely, 2011; Kuziemko et al., 2015).

In short, there are a number of rationales for governments to engage in redistributive activities. A key empirical question is how to best design these programs and test if they accomplish their goals. In doing so, the first aspect to consider is how to identify the poor to direct resources toward them (“targeting”). In [Section 2.1](#), we first discuss the potential strengths and weaknesses of common targeting methodologies, as well as describe the key factors that one must consider when planning randomized control trials (RCTs) on targeting methodologies. In [Section 2.2](#), we then discuss how to redistribute: for example, how big should the transfers be? How long should the transfers last? In the process of this discussion, we recount what we know and do not know from the current experimental literature. We also provide a guide for the design of RCTs to evaluate the types of social protection programs outlined earlier.

## 2.1 How to target the poor?

In high-income countries, targeting is often achieved by means testing: households bring a proof of income or unemployment to a benefits office, or they receive transfers through tax systems, such as through the United States Earned Income Tax Credit. However, in low-income countries, a lack of formal labor markets with a paper trail of income and employment status, coupled with underdeveloped tax systems, results in limited data to verify income.

To fill this data gap, low-income country’s governments can conduct income or consumption censuses. However, such censuses also present their own set of challenges, as anyone who has ever tried to conduct a survey module to elicit these kinds of data can attest: the modules take an inordinate amount of time and require certain skills, since one needs to map out all the different components of income (e.g., farming your own land one day, casual labor the next) or consumption (e.g., items purchased, the crops that one grows). Plus, without a formal mechanism with which to cross-check data, there is often nothing stopping surveyed populations from lying if they know that a cash prize is attached to their answers.

As such, low-income countries tend to develop alternative methods with which to identify the poor. The method chosen will depend on the priorities of the government: for example, Is the aim to target based on a particular poverty line?<sup>4</sup> Is it preferable to target

<sup>4</sup> Note that in addition to targeting poverty status, some programs also target particular demographic characteristics, such as whether a woman is pregnant or has children. As targeting on these characteristics tends to be easier—due to their verifiability—we will not discuss them in detail here for conciseness.

the poor based on income, consumption, or some other metric of poverty? How spatially dispersed are the poor? It will also depend on the institutions in place and context: how good is the implementing agency's ability to conduct surveys? How responsive are local leaders to citizens? We first outline key categories of targeting methodologies and then discuss how experimental methods can help distinguish between varying features of these methods.

Finally, note that we focus on targeting *poor* households in this section, given that the goal of redistributive programs is to provide impoverished households with a basic standard of living. Transfer programs that hope to enact longer run changes may target differently, depending again on their goals. For example, [Barrera-Osorio and Filmer \(2013\)](#) compare the effectiveness of scholarships when they are targeted to the poor versus when they are merit-based: while both increase school enrollment, only merit scholarships increase test scores. Thus, the method that you choose would depend upon whether you would like to redistribute scholarships to the poor, or to redistribute scholarships to those who will have the highest marginal return from them given a metric of test scores. We will revisit the idea of differences in targeting methods and goals later, when we discuss programs that aim to change long-run outcomes.

### **2.1.1 Targeting methods**

While there are many variations in practice, there are four primary categories that encompass most targeting methodologies:

- Geographic: If the poor are concentrated in particular villages, districts, or regions, giving everyone within those areas access to social protection may be an effective method to transfer resources to the poor (see, for example, [Baker and Grosh, 1994](#); [Elbers et al., 2007](#)). Moreover, this form of targeting may be particularly attractive when the institutional capacity needed to collect individual information is low, as only aggregate information is needed, e.g., poverty mappings, rainfall data, etc. Note, however, that this method may also be politically sensitive as it disburses benefits to some areas but not others.
- Proxy-means testing (PMT): In this method, the government collects demographic and asset data from households and uses these data to predict or “proxy” income or consumption.<sup>5</sup> Sometimes this method involves a quick-and-dirty poverty score card with just a few questions. Other times, a longer, more detailed asset and

<sup>5</sup> This is typically done with a nationally representative data set that includes the variable on which to target (e.g., income), as well as numerous demographic and asset variables. Next, income is regressed on different household characteristics, looping through different combinations and permutations of the variables, until the set of characteristics that best predicts income is identified (you find (often regional fixed effects or regional variables are also included for better precision). After conducting a census to obtain the chosen household characteristics, it is then possible to compute predicted income for each household using the formula. Households below a chosen cutoff of predicted income would thus be eligible for the program.

demographic survey is conducted. However, in either case, the key is to choose variables that are simple to collect, relatively easy to verify (e.g., whether the household has a dirt or concrete floor, or if they have telephone line), and that are less likely to be distortionary (e.g., school enrollment may predict poverty, but we may not want to incentivize households to keep their kids out of school). Households who pass the PMT—i.e., are below a certain poverty line—are then automatically enrolled. The effectiveness of the PMT will depend upon different factors: the formula’s predictive power, the quality of survey team, etc.

- Self-targeting: Self-targeted programs are those in which everyone is allowed to apply, but in which some sort of “barrier” is put into place to reduce the probability that rich households try to access the program. Theoretically, there are different barriers that can generate this form of selection ([Nichols and Zeckhauser, 1982](#)), ranging from time costs to apply, means testing on arrival ([Alatas et al., 2016](#)), and work requirements ([Besley and Coate, 1992; Ravallion, 1991](#)).<sup>6</sup> When done right, this can effectively screen out the rich (see, for example, [Alatas et al., 2016; Christian, 2014](#)). Given that households may fall into and out of poverty, these programs also have the potential advantage of flexibility, allowing households to access the programs when they have a bad shock. However, these programs also run a substantial risk: it may be hard for governments to initially predict, in advance, how many people will actually enroll or participate, providing additional challenges to budgeting and implementation.
- Community-based methods: In this method, community members choose who are needy in their locality. Theoretically, this method could not only bring in better local information on who is poor but it could also incorporate the community’s perceptions as to what determines poverty in their location ([Seabright, 1996](#)). A potential benefit is that this may make the program more politically popular, as people may feel that the list is more in-line with their vision of who is needy or deserving. A key worry is that by allowing for local discretion in choosing program recipients, elites may possibly capture the process ([Bardhan and Mookherjee, 2000](#)). Moreover, another potential downside is that while this method elicits better data on relative poverty within an area, it does not provide information across villages. For example, [Alatas et al. \(2012\)](#) show that the difference between the PMT’s and community’s ability to target based on consumption nearly doubles when the PMT is allowed to use its cross-village information.

Ultimately, the method and design will depend on a number of factors: the goals, context, institutional capacity, targeting budget, etc. For example, [Alatas et al. \(2012\)](#) show that the choice of community methods versus PMT can depend on whether one

<sup>6</sup> For certain types of in-kind goods—subsidized health products, insurance products—the price charged may also be used to select a particularly type of person who may value the particular product (see, for example, [Cohen and Dupas, 2010; Beaman et al., 2014](#)). We discuss this in more detail in [Section 4](#), when we discuss programs that are geared at changing longer-run income or behavior.

wants to specifically target a hard measure of poverty (e.g., income) or a soft one (e.g., perceptions). Similarly, choosing the right design of the PMT—e.g., the number of questions that goes into the formula—will also depend on the institutional capacity to administer the PMT. Moreover, [Beath et al. \(2013\)](#) show that whether aid allocations reach the neediest or not can depend on the type of community institutions that participate in targeting.

In practice, while there are distinct categories of methods, governments often mix and match the methods depending on circumstance. For example, they often try to save money by conducting a PMT on a selected sample of people who are likely poor, rather than conducting a full PMT census. For example, in its earlier form, Mexico's Progresa program conducted a PMT to determine eligibility only in the areas that were chosen as likely poor based on geographic targeting ([Schultz, 2004](#)). Similarly, Indonesia's Data Collection on Social Protection Programme (PPLS) uses community-based methods to help determine the list of households that will receive the PMT survey ([Alatas et al., 2012](#)). Kenya's Cash Transfer for Orphans and Vulnerable Children actually uses three methods: first, geographic targeting is used to determine locations, then community targeting is used in the selected areas to determine a list of households, and finally, those households are given a PMT to determine actual eligibility ([The Kenya CT-OVC Evaluation Team, 2012](#)).

### ***2.1.2 Experimentally testing between targeting methods***

Are experimental methods important for testing across targeting approaches? Not necessarily. For example, one simple research design would just be to try out two different methodologies in the same areas and then compare the income levels of those selected under both methods.

While this research strategy may prove attractive in some ways, it may miss out on the nuances of targeting that may ultimately be quite important in understanding the relative effectiveness of the differing methods. First, many of these targeting methods require considerable effort on the part of citizens and the program staff.<sup>7</sup> Simulating conditions to be as real as possible, with at least a small amount of cash on the line for the households that are to be selected, is important to understand how the methods would actually work in practice.<sup>8</sup> In terms of citizens, people might behave differently during the process if they do not believe that the targeting list will actually be used to distribute resources.

<sup>7</sup> One obvious exception to this is geographic targeting, which does not rely on field operations, such as household and community leader interactions. If a different method was conducted for the actual program, then simulating geographic targeting over the same areas based on administrative data can provide an accurate comparison of the type of person selected under both methods. However, by not conducting geographic targeting in practice, it will not be feasible to test program outcomes other than just who is selected (e.g., political acceptability, leakages) that result from using different targeting methods.

<sup>8</sup> You could, for example, try out two methods in the same area and offer a transfer to everyone who is selected by either of the different methods, but then staff or villages may coordinate so that different people are on each list. And, it runs the risk of confusing people, so that they do not take the exercise very seriously.

For example, they may not exert any effort in discussing and ranking households in community targeting, or individuals under self-targeting may just not bother to show up even though they would, if there were real cash involved. Moreover, people may give more truthful answers in the PMT, or not claim a greater poverty status than their reality at a community meeting, if they believe that their answers do not have any consequence. In terms of staff, the results may noticeably differ if you test out methods without stakes with a highly trained set of staff than if trying out the methods with the typical type of staff that would be hired and trained ([Alatas et al., 2012](#)).

Second, the choice of targeting method may affect both the program and household outcomes. For example, the targeting method chosen may help determine the ultimate satisfaction of the program, which in turn may affect program acceptance and the government's ability to implement the program. Targeted programs can be controversial, with some households receiving a transfer, while their neighbors do not. If citizens believe that a certain method is unfair and that it would produce a flawed list, they may be less likely to support the overall social protection program and potentially block the distribution of benefits.<sup>9</sup> Moreover, if people believe that the wrong individuals were chosen due to the fact that a certain method was used, it may possibly lead to distortions in how informal insurance or lending operates within a community.

Finally, who is chosen may be different than who actually receives the transfer ([Alatas et al., 2013](#)) and this may also vary by the targeting method that was employed. For example, suppose that the PMT better selected the poor than a community method. But that the PMT had less legitimacy than the community method, so that village leaders did not adhere to the PMT in practice when distributing the transfers, but they did adhere to the community list. In this case, simply simulating both methods in the same area to elicit a beneficiary list would possibly wrongly suggest that one should select the PMT since you would only be able to study who was chosen but not be able to measure who ultimately would get the benefits.

In short, for all of these reasons, we would want to randomize the targeting methods to different areas to see how the method affects who is chosen and who ultimately receives the transfers. The optimal design for a field experiment in the domain of targeting will depend not only on a number of factors, including which method is being studied, but also what the particular context looks like. However, there are a number of key questions to keep in mind regardless of the given design.

The first question is whether or not it is necessary to have a control group, in the traditional sense. RCTs generally compare the outcomes from a treatment group that

<sup>9</sup> For example, [Alatas et al. \(2012\)](#) show that community targeting led to much higher levels of satisfaction than the PMT, with village leaders feeling less comfortable making the transfers under the guise of public scrutiny when the PMT had been used. They provide suggestive evidence that this difference is due to the perceptions of the methods, and not the ultimate lists that the different methods produced.

receives an intervention with those of a control group that does not. In this case, since the outcome of interest is who is selected under different targeting methods within the same program, it may be viable to simply have multiple treatment groups where each is randomly assigned a different targeting method, but everyone receives the program.

Second, at what level should the randomization take place? Randomizing at the individual level offers the most statistical power, but in this case, the targeting treatments often involve some sort of group participation (e.g., community, self-targeting) or group data (e.g., geographic). Moreover, even with a PMT, where it is possible to vary how the survey is conducted across individuals, a question of interest might also be whether the program administrators change how they respond to the targeting list in their area based on which method generated the list. Thus, in most cases, it is appropriate to randomize across a sensible geographic unit; as such, power calculations to determine sample size should account for the group structure of the data.

Third, is a baseline survey needed? With most experiments, the answer is not necessarily: the analysis will consist of comparing outcomes of those in the treatment and the control group. A baseline might help for power if the outcome measures within a person are highly correlated across time and it may allow you to test for the heterogeneous treatment effects by various baseline characteristics (Duflo et al., 2006). But, it is not necessary per se in a typical randomized experiment. In this case, a baseline is essential: a key outcome of any targeting experiment will be the baseline income or consumption levels—prior to the targeting—of those who are actually chosen.

Fourth, what kinds of data should be collected? The exact variables would, of course, depend on what methods are being tested and what outcomes are expected. But, typically, in the baseline, it is essential that data be collected on the variables being targeted on (e.g., income, consumption, etc.), so that inclusion and exclusion errors can be computed.<sup>10</sup> To measure distortions in who is chosen along certain dimensions, it may be worthwhile to collect baseline data on political affiliations and relations to political leaders. In the end line surveys, valuable data to collect could include who actually received the program, satisfaction levels, and metrics on general program functioning.

Finally, will the experiment just aim to measure the reduced form effect or will it also attempt to understand why it is working (or not)? If the only relevant question is comparing method one versus method two (i.e., the reduced form difference of the two programs), only two treatment arms are needed. But we know that the effectiveness of a method may vary based on its own design, and therefore, it may be worthwhile to learn more about a method's effectiveness if certain details of the implementation are varied. Here, theory can help guide the appropriate subrandomizations: For example,

<sup>10</sup> Alatas et al. (2012) also ask villagers to rank one another to gain the “average” person’s belief about another household’s poverty status in a village, as well as ask household to assess their own poverty status.

[Alatas et al. \(2016\)](#) experimentally compare the outcomes of a PMT with a self-targeting mechanism within Indonesia's conditional cash transfer (CCT) program. Importantly, they experimentally vary the distance of the application site under self-targeting to generate exogenous variation in the cost of the application "barrier." They then use this variation to estimate a model of the decision to apply for the program and simulate self-targeting outcomes under different levels and types of application costs.

## 2.2 Evaluating the impacts of redistributive programs

Once the poor have been identified, the next task at hand is to think about whether the redistributive programs are achieving their goals in practice. The simplest redistributive programs are those that entitle the identified poor to some form of cash stipend to provide a certain standard of living and are *unconditional* on any behaviors (an unconditional cash transfer, or UCT). The Chinese Di-Bao Program is the largest UCT in the developing world, reaching 78 million households (see [Chen et al. \(2006\)](#) for a description of the program). Other prominent examples include South Africa's Child Support Grant, India's National Old Age Pension Scheme, and Kenya's Hunger Safety Net Program.

Evaluating these programs usually follows a typical design. First, poor households are targeted in the sample area. Then, potential beneficiaries are randomly assigned to the treatment group ("receives transfers") and the control group ("does not receive transfers") to assess the program impacts. Examples of unconditional cash programs that have been evaluated in this fashion include the Zambia Child grant program ([Jessee et al., 2013](#)), the Kenya Hunger Safety Net program ([Merttens et al., 2013](#)), Kenya's Cash Transfer for Orphans and Vulnerable Children (see, for example, [The Kenya CT-OVC Evaluation Team \(2012\)](#) and [Covarrubias et al. \(2012\)](#), on the Malawi Social Cash Transfer Scheme).<sup>11</sup>

Even though the overall evaluation strategy is relatively straightforward, there are still a number of decisions to be made—concerning the level of randomization, the type of data collected, the timing of data collection, etc.,—that are important in assessing impacts.

*The level and form of randomization:* Transfers are generally provided to an individual or household, and so it is tempting to randomize at the individual level to maximize statistical power. For example, [Schady et al. \(2008\)](#) do exactly this. However, [Angelucci and de Giorgi \(2009\)](#), among others, show that there may be spillovers of transfers from

<sup>11</sup> There were two RCTs conducted on *Bono de Desarrollo Humano*—one on child health ([Paxson and Schady, 2010; Fernald and Hidrobo, 2011](#)) and one on education ([Schady et al., 2008; Edmonds and Schady, 2012](#)). The program was initially supposed to be conducted as a conditional cash transfer program and some announcements were made to this effect ([Paxson and Schady \(2010\)](#) had multiple treatment arms to test between pure cash and cash with conditions), but the conditions were never enforced. Given that just the framing of the program as a CCT could still have impacts, we do not include this in our discussion of UCTs but instead discuss it later.

eligible to ineligible households within a village.<sup>12</sup> This implies that the randomization likely needs to be done at a higher level, at the village level or subdistrict level depending upon what types of spillovers one might expect.<sup>13</sup>

If randomization is at a group level, it is vital to have enough “units” to randomize over or it will be challenging to measure impacts. For example, the Kenya’s Cash Transfer for Orphans and Vulnerable Children ([The Kenya CT-OVC Evaluation Team, 2012](#)) had only 28 units of randomization, which might account for their difficulty in detecting any program impacts; the Malawi Social Cash Transfer Scheme had only 8 clusters and did not fully account for the grouped nature of the data in the analysis ([Covarrubias et al., 2012](#)).<sup>14</sup> Thus, one should determine in advance what the desired size of treatment effects is (i.e., large enough for the program to be cost-effective, etc.) and use this to assess the statistical power of the proposed design.

Importantly, it is also key to identify the beneficiaries in the control group, not just the treatment group. Suppose a randomization determined which villages obtained a UCT and which were in the control group. In the UCT villages, targeting would have first been conducted to choose the beneficiaries within the village. However, unless a similar targeting strategy was also conducted in advance for the control group, it would be difficult to know who were the hypothetical beneficiaries in the control group (see, for example, [Covarrubias et al., 2012](#)). In this case, it would be possible to estimate the impact on the entire village as a whole but not easy to estimate the impact on just the beneficiaries.<sup>15</sup> Thus, it is essential, when possible, to use the same targeting methods in the treatment and control group, even if the control group is not receiving the program at the time of the study.

*Data collection:* What data should be collected and when? If the goal is simply redistributive, we may simply care whether poor households were actually identified and whether or not they were receiving their entitlements. In many developing countries, due to weaker institutional structures, corruption, or imperfect information on their entitlements, households do not receive the full transfer. So, an important set of survey questions should be focused on whether households actually received the transfer, whether they received the full transfer, whether the village elites held them up for a portion of

<sup>12</sup> As we discuss later, [Angelucci and de Giorgi \(2009\)](#) evaluate a conditional cash transfer program, but the basic ideas on spillovers hold for UCTs as well.

<sup>13</sup> As we discuss later, one might even want to design the study to capture different types of spillovers and general equilibrium effects.

<sup>14</sup> In the analysis of these cases, it would be necessary to adjust standard errors to account for both the grouped nature of the randomization and the small number of groups in keeping with the procedure outlined by [Cameron and Miller \(2010\)](#).

<sup>15</sup> Of course, this is not a problem within the village if the program is geographically targeted and everyone in all sample villages is eligible. However, this may also generate spillovers over larger geographic regions that one may want to be aware of, e.g., if everyone in a district or province receives additional income, would this increase demand for food, raising food prices?

their transfer, and so forth. Next, one may want to administer a household income or consumption model to measure household's income status.

If we simply care about redistribution, we would not necessarily care about how households spend the transfers—we would just care about whether they receive the money. Thus, if motivated by such a philosophy, one may argue to only measure the transfers and not bother examining what happens with the money transferred. But for many reasons, researchers do collect more data. For example, one political rationale of many redistributive programs is to provide a basic standard of living for children, so one may want to collect measures on child health, nutrition, and education. There are a number of issues to consider in doing this type of larger data collection, but we will revisit it later when we discuss broader issues surrounding behavioral change for program beneficiaries.

Next, a key question is whether or not a baseline survey is needed. In assessing the impact of a transfer programs, a baseline survey is not necessarily needed, since treatment was assigned randomly, meaning that a postintervention of the two groups is an unbiased estimate of the true impact. However, there are two key exceptions to this. First, if we want to assess whether the poor were indeed targeted and whether the poorest of the poor were able to receive their entitlements, then we would need baseline consumption or income data. Second, if particularly vulnerable subgroups are of importance (e.g., the very poor, families with children who are less likely to attend school, etc.), one may want to collect data on these groups—if administrative data of this sort is unavailable—to be able to stratify the randomization.

Finally, when do you conduct the follow-up surveys? Again, it depends a bit on the research aims. A follow-up survey should generally be conducted within the duration of the program to understand whether households are receiving the transfer. However, as we discuss later, if the longer-run impacts of transfer programs are of interest, one may also want to do additional surveys sometime after a household is no longer enrolled in the program.

### **3. MISSING INSURANCE MARKETS**

The poor face many risks, such as unexpected health costs, agricultural damage, or job losses. Without fully functioning financial markets, households may be unable to borrow to smooth consumption. Furthermore, even if households do smooth consumption in the face of shocks, they may be doing so in the short run by making longer-term sacrifices, such as pulling children out of school or not investing in health. At the extreme, natural disasters—e.g., earthquakes, flooding, famine—may cause group shocks that further limit the functioning ability of informal insurance mechanisms within a region.

Thus, an important role of social protection may be to help alleviate the impact of such shocks by providing social insurance of various forms. Two prominent forms of

insurance are agricultural insurance and health insurance, but we will not cover them here since they are discussed elsewhere in this handbook. Rather we will briefly discuss two other forms of insurance: disaster relief and unemployment insurance (UI).

Disaster relief is an important form of social insurance. While some quasi-experimental work has been done to study both the consequences of a disaster, and the distribution of aid, to our knowledge, there has been less experimental evidence in evaluating social protection in humanitarian settings. Notable exceptions include [de Mel et al. \(2008\)](#), who studied the effect of cash transfers to microenterprises in posttsunami Sri Lanka, as well as [Aker \(2014\)](#) and [Hidrobo et al. \(2014\)](#) who examined cash versus other types of transfer programs in informal refugee camps in the Democratic Republic of Congo and Northern Ecuador, respectively. Part of the reason for the lack of experimental evidence comes from ethical concerns with a need to provide immediate assistance trumping research and evaluation. Related is a logistical challenge: the chaos and highly transient nature of refugee camps can make randomization particularly challenging to implement. However, given the about 60 million refugees and internally displaced people worldwide ([UNHCR, 2014](#)), understanding how safety net programs can be better tailored to provide assistance in humanitarian settings has been an increasingly urgent need, particularly as many refugee camps persist for long after the immediacy of the disaster.

A second important form of insurance is UI, the provision of temporary assistance to those out of work. This form of social protection has been traditionally missing from poor countries, as the data-poor environments preclude easy identification of those who are working in full-time positions to then determine who is temporarily out of work. But, it is becoming more common in relatively more developed, low-income countries with formal labor markets (e.g., Brazil, Egypt). While there have been numerous experiments that have tried to understand aspects of UI in developed countries, to our knowledge, there is little experimental evidence on UI in lower to middle income settings.<sup>16</sup> One notable exception is [Mickelwright and Nagy \(2010\)](#), which randomized unemployment beneficiaries in Hungary to varying levels of monitoring (i.e., visiting the employment office every 3 months with no job search questions versus visiting every 3 weeks to answer questions on job search behavior). However, as UI spreads and becomes a more important component of social protection in developing countries, empirical evidence will be needed to understand its impacts on labor markets and household outcomes. Empirical evidence will also be needed to understand how to better design these programs—addressing how to best verify employment status, how to best distribute benefits, and how to ensure that the programs provide incentives to find work.

<sup>16</sup> For example, [Meyer \(1995\)](#) summarizes a number of early experiments on UI in the United States, focusing on four cash bonus experiments and six job search experiments. More recent examples include [Van den Berg and Van der Klaauw \(2006\)](#), who explored the effect of additional monitoring and counseling on UI recipients in the Netherlands; [Grenier and Pattanayak \(2011\)](#), who measured the impact of legal assistance on UI in the United States.

## 4. BEHAVIORAL CONSTRAINTS

It is often argued that behavioral or household-bargaining constraints influence the choices of low-income households, thereby perpetuating poverty. Regardless of whether or not these behavioral constraints are present, safety net programs are often designed to correct these types of constraints and encourage socially optimal behavior.

There are two key forms that behaviorally focused social protection programs usually take. The first is providing in-kind transfers rather than cash, under assumptions such as people will be too tempted to spend money on “bad” goods (such as tobacco and alcohol) or that household-bargaining constraints would imply that cash transfers would cause socially undesirable outcomes, such as an underinvestment in children or an exit from the labor market.

The second is to directly condition the receipt of transfers to households on a household’s compliance with certain long-term investments, such as the children attending school or visiting health clinics. These CCT programs have become increasingly common in Latin America and have begun to spread to other parts of the world, existing in more than 52 developing countries as of 2013 ([Fiszbein and Schady, 2009](#); [Saavedra and Garcia, 2012](#); [Gentilini et al., 2014](#)). And, finally, note that some programs do a combination of both in-kind and conditions, with 130 countries doing some sort of conditional, in-kind program.

Careful experimentation that helps test the underlying rationale for these programs can help determine if behavioral constraints exist, and if so, what form they take. Thus, they can provide insights into whether these kind of constraints on behavior actually improve welfare or simply make redistributive programs more expensive.

### 4.1 Providing in-kind transfers to constrain spending choices

#### 4.1.1 *The rationale for in-kind programs*

Unconditional, in-kind redistributive programs typically provide free or highly subsidized goods to program recipients. They can entail a direct provision of the goods, such as a food or fuel transfer program; the subsidy of a product distributed through local governments, NGOs, or designated shops; or a system of vouchers that are constrained to particular types of goods, such as a food stamp program.<sup>17</sup> [Gentilini et al. \(2014\)](#) note that 89 low-income countries have unconditional, in-kind transfer programs, with examples

<sup>17</sup> It is important to note that there are other types of in-kind transfers, such as ones that provide free health products ([Cohen and Dupas, 2010](#); [Dupas et al., 2013](#); [Ma et al., 2013](#); [Glewwe et al., 2014](#)), prizes or scholarships for school ([Berry, 2014](#); [Kremer et al., 2009](#)), school meals programs ([Kazianga et al., 2012](#); [Vermeersch and Kremer, 2004](#)), and public housing ([Kling et al., 2004](#)). These programs may also have different aims, such as solving the externality issue in health product take-up. As they are discussed in other chapters in this handbook, we refrain from discussing them here.

including Indonesia's Rice Subsidy Program ("Raskin") and the Public Distribution Systems in both Bangladesh and India.

There is much debate about whether transfers should be in-kind. If you ask a typical economist, most will favor cash programs, under the idea that households will maximize utility if they have choice over what they purchase, rather than receiving a good of equal monetary worth that they may not value as much. However, there are a number of arguments proposed in favor of in-kind subsidies (see [Curie and Gahvari \(2008\)](#) for an excellent review), which may explain their general persistence worldwide.

The most cited explanation is that of paternalism: People often have an image of the lazy, out-of-work husband co-opting the family cash and wasting it on alcohol, tobacco, and other forms of entertainment, rather than making spending decisions that can improve the family's living situation or investing in children. Thus, the argument follows that an in-kind subsidy could reduce the husband's ability to do so, forcing redistribution within the household to the women and children who typically have less household-bargaining power. However, others argue that if the in-kind subsidy is inframarginal—or if it is easy to resell goods—then in-kind subsidies will not alter the household's consumption bundle, and it is simply a more costly mechanism to redistribute to the poor than cash. The experimental evidence thus far suggests that cash programs do not generate more spending on temptation goods, such as alcohol and tobacco. Nonetheless, in-kind programs are often "sold" as targeted to women and children and thus tend to be more popular among tax payers who want to ensure that their tax dollars are not wasted, but rather used to "feed children" and reduce social ills such as school dropouts and crime (for example, see [de Janvry et al., 1991](#); [Epple and Romano, 2008](#)).

A second potential reason to favor in-kind transfers over cash transfers is that they might have lower de-incentive effects on work and may in fact spur work. A common worry with cash transfers is that they could provide a disincentive to work, particularly if households worry about losing their benefits as their income rises above the eligibility line. It has been argued that in-kind transfer generates fewer labor market distortions and may in fact be labor market enhancing if the provided good is a complement to work. For example, in areas where productivity is low due to nutritional constraints, a food transfer program could ease this constraint. However, the existing evidence thus far does not imply that cash transfers greatly reduce labor market participation (for example, see [Alzúa et al., 2013](#); [Banerjee et al., 2015a,b,c](#)), perhaps due to the long duration of benefits and uncertain processes for recertification observed in many developing countries. Furthermore, the cash transfers may also help ease credit constraints for those engaged in the agricultural sector, increasing the productivity of agricultural labor ([Gertler et al., 2012](#)).

A third reason in support of in-kind programs is their self-targeting properties ([Besley and Coate, 1991](#); [Christian, 2014](#)): by providing a good that the poor differentially value relative to the rich, the poor will apply and the rich will opt out. Again, if money is fully

fungible, richer households could simply opt in and sell these goods for the cash, etc. However, we might expect that transition costs and other constraints would imply that in-kind goods are viewed differently than just pure cash. Note that despite this rationale, many in-kind transfer programs are independently targeted prior to distribution, shutting off a channel through which an in-kind program may generate these types of effects. And, while the nonexperimental evidence suggests that in-kind programs are better at selecting the poor (see, for example, [Jacoby \(1997\)](#)), there is little experimental work that compares the magnitude of its targeting properties against different forms of means-tested cash programs.

A final argument in favor of in-kind transfers comes from the idea of missing markets, i.e., if for some reason the market does not provide the good on its own, just providing cash may not be enough. This may be a particular issue in remote areas where high transport costs discourage the spread of products or in disaster or war zones, where food and other products may be in short supply.

#### **4.1.2 Evaluating in-kind programs**

One can evaluate an in-kind transfer program in a manner similar to evaluating a cash program, i.e., randomize some areas to receive in-kind transfers (treatment group) and others to not (control group). Then one can collect data to understand if poor households were indeed properly targeted and whether or not they were able to access their entitlement or subsidized products.

However, unlike pure redistributive programs, one may also want to understand whether the theoretical behavioral constraint actually exists, as well as whether the behavioral change that one aimed to induce has occurred. As such, there are two other design features to consider. First, while an evaluation of a redistributive program would be focused on collecting variables to determine whether or not the redistribution has successfully occurred, an evaluation of an in-kind program would additionally require collecting variables to test for the hypothesized behavioral changes.

For example, in evaluating a food transfer program, if the goal is to increase food consumption for children, one would care about collecting detailed modules on food and calories consumed, as well as health indicators for children. However, note two important caveats of this data collection process. First, as [Schady et al. \(2008\)](#) point out, people may hesitate to provide accurate information on surveys if they believe that the surveys are connected to reverification for the program. In which case, one may want to collect variables that are easier to verify: Assets that one can observe, vaccination records on a card, body mass index, other anthropometric variables to assess the health status of children, etc. Second, as the transfer program may be designed to address a number of behavioral changes, one's worry is that in collecting many different variables, we would find some significant impacts just by chance. Thus, one might want to prespecify some of the key hypotheses and outcome variables in advance ([Miguel et al., 2014](#); [Olken, 2015](#)).

Second, and perhaps more important, one may also want to consider evaluation strategies that isolate the behavioral constraint. The basic RCT described earlier comparing the effect of a specific program against control areas that do not receive any assistance is important for understanding total program effects, but it does not tell us how the particular constraint (e.g., providing rice versus cash) affects the observed outcomes. Understanding the relevance of specific behavioral constraints can be important, if they imply specific changes to program design.

Therefore, comparing in-kind transfer programs to cash transfer programs can provide useful insights into whether or not constraining behavior is welfare improving. For example, [Aker \(2014\)](#) explores the effect of cash versus food vouchers for displaced households living in an informal camp in the Congo and shows that the level of food consumption was the same regardless of the mechanism since voucher households simply bought food that is relatively easy to sell (e.g., salt). Similarly, the Mexican government took advantage of multiple treatment arms to measure both the effect of in-kind programs to cash and to doing nothing at all. Specifically, they compared three treatments: (1) an in-kind transfer program that gave households 10 different items of food, (2) a cash transfer intended to be of equivalent value, and (3) a control group that did not receive transfers.<sup>18</sup> While the in-kind program distorted consumption of some individual types of food ([Cunha, 2014](#)), overall food consumption was similar across both programs ([Skoufias et al., 2013](#)) and there was no difference in observed weight among women ([Leroy et al., 2013](#)).

Other studies have also collected valuable data that highlight the trade-offs between programs that may satisfy a society's specific goal that an in-kind program may be designed to address (e.g., greater food consumption for kids) versus other important social goals. [Hidrobo et al. \(2014\)](#) compare cash, food, food vouchers, and a control group in Ecuador and find that all three types of programs improve per capita food consumption and caloric intake. However, while food and food vouchers increase calories and diet diversity a bit more relative to cash, the cash program is much easier and cheaper to implement, which may be of real concern for countries with weaker institutional quality.<sup>19</sup>

Similarly, [Hoddinott et al. \(2014\)](#) compare cash versus food transfers in Niger and find that the food transfer program had a larger effect on food consumption and diet variety than cash. However, households were not wasting the funds on temptation goods such as alcohol; they used cash to invest in greater agricultural inputs. Thus, using their estimates,

<sup>18</sup> Note that a particular challenge is ensuring that the in-kind transfer is equivalent to the cash transfer. For example, in the PAL program, the value of the food transfer was 30% more than the cash treatment, since the food basket was based on wholesale prices to the government rather than the prices that consumer pay (see Cunha, 2014 for a discussion of how to make them ex post equivalent).

<sup>19</sup> In an interesting follow-up paper to this experiment, [Hidrobo et al. \(2013\)](#) show that transfers (irrespective of whether food, cash, or vouchers) reduce intimate partner violence.

one can think about how to model the trade-off between the additional utility households receive from spending as they choose, relative to society's utility from the shift in food consumption.

In short, the theory suggested that food, cash, and vouchers could have different effects on household outcomes, but the existing evidence mostly shows that this did not materialize in practice—likely because the particular items of food distributed in these contexts were inframarginal.

## 4.2 Adding conditions to incentivize behavior

Since the 1990s, it has been increasingly common for many developing country transfer programs to layer a level of “conditionality” onto redistributive transfer programs, e.g., requiring children to go to school and get health checkups for the family to receive its full transfer. The conditions usually aim to correct an underinvestment in the family’s well-being that stems from some sort of market failure within the household—parents not internalizing their children’s full returns to school, the family not internalizing the benefits of vaccination for society, etc.

CCT programs worldwide have been studied by RCTs—from Mexico’s Progresita/Oportunidades to the Philippines’ Pantawid Pamilyang Pilipino Program (PPP)—showing important effects. As with the other programs we have discussed, one can evaluate CCTs by simply randomizing areas to receive or not receive the CCT and then studying the outcomes; this tells us the reduced form effect of having the program relative to no program. In doing so, many of the issues that we have raised earlier—what data to collect, when and how to target, etc.,—would thus be important to think about in this context as well.

However, given the unique nature of the CCT in attempting to correct an investment failure within the family—in addition to redistributing to poor households—more sophisticated analysis can be done to better understand the role of the conditions on household behavior. Later, we discuss three key questions that experimentation can help shed light on (1) What is the impact of the conditions relative to basic redistributive programs?; (2) What should the conditions be?; and finally, (3) How do we enforce them?

### 4.2.1 Evaluating conditions relative to basic redistributive programs

Evaluations that test a conditional transfer versus no program have difficulty in teasing apart the conditionality mechanism from the liquidity or income effect of the transfer itself. Whether to include conditions in a cash transfer program is an essential design decision. Conditions require a budget to fund them and staff to enforce them. They may also generate selection effects on who chooses to participate.

To isolate the effects of the conditionality, a simple experimental design would have one treatment with a CCT, a second treatment with UCTs, and a control. Thus the

second treatment group generates a pure income effect, which allows one to learn whether the conditionality changes behavior, or whether the CCT changes behavior simply through its income effect.

This is the approach taken by [Baird et al. \(2011\)](#) in the Zomba Cash Transfer Program in Malawi. The CCTs performed better at inducing the desired behavior: School dropout rates were lower in the CCT arm than the UCT arm and persisted beyond the program's end. Moreover, attendance rates improved and cognitive ability, mathematics, and English reading comprehension test scores increased for the CCT arm but not the UCT arm. Thus, the conditions had effects on households above and beyond just providing cash. However, the UCT arm had significantly lower marriage and pregnancy rates. This initially may feel like a counterintuitive result since schooling is thought to postpone marriage and lower pregnancy rates. However, the UCT is provided to households even if the teenage girls do not attend school, whereas only those who attend school receive the CCT. Thus, the effect on marriage will depend largely on the relative size of two groups, the group of households that does not attend school under either the UCT or CCT, compared to the group of households that attends school under the CCT but not the UCT. If the income effect on marriage and pregnancy in the first group is large enough ([Duflo et al., 2010](#); [Ferre, 2009](#); [Osili and Long, 2008](#); [Ozier, 2011](#)), the UCT will lead to, on net, lower marriage and pregnancy rates. The authors of the Malawi study found this to be true. This dynamic illustrates that the conditions themselves can reduce redistributive aid to the poor—and any associated benefits of the aid—by changing who participates in the program. And, it reinforces the need for careful data collection on ancillary outcomes—e.g., in this case, marriage and pregnancy—to understand the full set of mechanisms through which a program works so that policy-makers can better understand the trade-offs that they would make by implementing one program over another.

Further insights on the trade-off between a UCT and CCT can come through examining what beneficiaries would choose, given a choice. Most programs naturally do not provide such a choice, but providing a choice in an experimental context can help offer insights into which programs would yield higher utility to citizens. For example, individuals could use the conditions in the CCT to generate a personal or family commitment device to engage in future behavior. If respondents were to opt in to such a CCT over a UCT, this is strong evidence of either individual demand for a commitment device ([Ashraf et al., 2006](#); [Bryan et al., 2010](#)) or of demand due to family conflict over education or health decisions. Similarly, in Brazil, researchers examined households that were given a choice between a UCT and CCT on school attendance and also tested a subtreatment within the UCT in which households were informed of whether the children were attending school or not ([Bursztyn and Coffman, 2012](#)). The parents exhibited a strong preference for the CCT, unless the UCT included monitoring of their children's school attendance, in which case they were content with the UCT. These preferences lend an

important insight into the underlying mechanisms of the CCT and suggest that the conditionality in this context was simply a tool for parents to better monitor children. Thus, interventions that improve monitoring and communications between schools and parents may be a better solution than the complicated nature of CCTs over UCTs.

### **4.2.2 Which conditions should be imposed?**

Policy-makers have to make choices about what conditions to impose: Adding conditions adds additional costs of monitoring and as we discussed earlier can have important implications on who participates in programs, potentially screening out the very people whom you want to reach. Thus, experimentation can also help along this dimension, helping to determine which conditions are more impactful and worth focusing on.

In doing so, it is more complicated than just choosing to condition on a single factor, say “schooling.” The structure is also important: Conditions can be imposed on inputs or activities (e.g., attendance of children) or outcomes (e.g., test scores) or both. And then, the payment structure must also be defined. For example, [Barrera-Osorio et al. \(2011\)](#) show that whether you simply condition a monthly payment by attendance, or hold back part of the payment and provide it only if the child reenrolls in school, can affect schooling outcomes.

Another type of structure provides conditions for participation but does not financially penalize households if they do not meet them. Thus, this structure essentially provides households with a “nudge.” For example, comparing a CCT with a cash transfer program that was “labeled” for schooling (LCT), [Benhassine et al. \(2015\)](#) show that both types of programs improved school participation relative to the control group, but they were not statistically different from each other. The conditionality costs more to implement and so was more expensive relative to the LCT.

As we discussed earlier, applying conditions may lead to a trade-off between the goals of the conditions and the initial goal of redistribution. One legitimate worry is that certain types of conditions may discourage poor households from applying since they find some of conditions too onerous to comply with, thus diminishing the ultimate redistributive goal of these programs. For example, in areas with strong cultural beliefs against vaccines, would requiring vaccines for children reduce the probability that the poorest households participate in the program? One can imagine extensions of the CCT literature with randomization not only for whether the program has conditions but also for the types of conditions in different areas—measuring not only the effect on recipients but also the effect on who applies for the programs and how well the implementers can realistically enforce them.

### **4.2.3 Enforcing the conditions**

The enforcement of the conditionality is critical to examine. Without the enforcement, the program is perhaps theoretically equivalent to an UCT. Politics at the policy level and

corruption at the implementation level, both can lead to de facto removal of the conditionality. This occurred, for example, in Ecuador with the *Bono de Desarrollo Humano (BDH)* program. An evaluation of the program thus analyzes the results as if the program is an UCT ([Fernald and Hidrobo, 2011](#); [Paxson and Schady, 2010](#)). However, a CCT program that fails to enforce may ultimately generate a different behavioral response than just a pure UCT. For example, in the Ecuador case, many households believed that the funds were indeed conditional, even though in reality they were not.

Enforcement of conditions is an area worthy of further research, but it is not obvious that this is an appropriate randomized trial territory. While one could randomize the enforcement of the program, in an approach similar to what [Olken \(2007\)](#) uses for road building, the political environment that led a government to fail to enforce the conditions may matter, and it may be that merely randomizing enforcement in one setting does provide insights to settings in which enforcement is not viable for political or social reasons. The reason for lack of enforcement is important and cannot be simulated merely through randomization. For example, lack of viability of enforcement could be driven by constituent expectations, local social norms (which both drive the level of enforcement and the treatment effect of the program), or simultaneous policies that interact with the treatment effects of the transfer program.

However, an unenforced condition may be similar to a suggestion or a “nudge.” For example, an unenforced condition of school attendance may serve a similar role as the government merely labeling a transfer as an “education support program.” [Benhassine et al. \(2015\)](#) examine this question by designing a two-pronged experiment: CCT versus LCT (those two prongs were crossed with a household structure test, providing the program to mothers versus fathers). To understand the underlying mechanisms, the study collected data on process changes and also used a multiarmed experimental design. For example, to understand if the conditionality (or labeling) leads to changes by signaling information about returns to education, researchers collected data on parental beliefs about returns to education (no change was found). Attendance (conditional on enrollment) increased after CCTs and LCTs, leading to an increase in time spent on studying, at school, and traveling to school and a decrease in leisure and productive labor (but not a decrease in chores); this suggests the barrier to schooling was due more to a lack of student interest (thus drawing a similar conclusion on mechanisms as the Brazilian study mentioned earlier ([Bursztyn and Coffman, 2012](#))). The multiarm experimental design then tested explicitly the marginal benefit of the condition, over and beyond a merely labeled transfer, and the study found no additive effect beyond the label.

## **5. MARKET FAILURES PREVENTING ASSET ACCUMULATION**

Most of the issues we discussed thus far focus on providing short-term relief through redistributive aid or insurance to households, with some programs paying attention to

ensuring that the money is spent on goods and services that benefit households in ways beyond an increase in cash or liquidity. However, certain forms of antipoverty programs also aim to address underlying mechanisms that may be creating poverty traps to improve long-term income for the poor or to provide mechanisms for individuals to build long-term financial assets for when they are elderly. These are, indeed, more complicated challenges: If the underlying market frictions are beyond credit and savings market constraints, then the solution will require more than redistribution. If redistribution alone is employed, it may provide important short-run benefits but may ultimately act as more of a band aid on immediate symptoms without helping individuals achieve a sustained increase in income.

## 5.1 Building productive assets

For the past 30 years, microcredit has been a leading development policy in the fight to reduce poverty. Unfortunately, seven recent randomized trials have shown that microcredit,<sup>20</sup> while it does provide important benefits, does not improve long-term income, on average, for participants in its current form (for a review, see [Banerjee, Karlan, and Zinman, 2015](#); the seven randomized trials are [Angelucci et al., 2015](#); [Attanasio et al., 2015](#); [Augsburg et al., 2015](#); [Banerjee et al., 2015a,b,c](#); [Crépon et al., 2015](#); [Karlan and Zinman, 2011](#); [Tarozzi et al., 2015](#)). In addition, [Meager \(2015\)](#) aggregates the microdata across these studies and uses Bayesian hierarchical models to show that the effect of microcredit on household profits is likely very small and that the effects from each individual study site are reasonably informative for each other. This suggests that either credit constraints are not driving stagnated growth for the poor, the current designs of microcredit programs do not fully address credit constraints, or microcredit may not work without changes in other conditions that also generate market failures for the poor. Multisite studies provide tremendous opportunities for such analysis, e.g., through building more robust theories and then examining, using appropriate statistical tools, how well results from multiple sites fit broad theoretical frameworks.

Indeed, the poor do face multiple failures at the same time that may hinder long-run investment and income growth. But we typically observe programs tackling one problem at a time, rather than a coordinated system. Poverty alleviation policy, as with many government programs, often operates in silos. One silo, described previously and typically managed under the umbrella of “social protection,” focuses on redistribution policies through either CCTs or UCTs. A second silo, often managed under the ministries of trade or agriculture, focuses on livelihood support, such as the transfer or a productive asset or agricultural input, alongside some training. A third silo, financial inclusion, is

<sup>20</sup> Note that this lesson was only learned by running similar experiments across different project sites and countries, showing how valuable replication studies can be in changing perspective.

often administered through for-profit or nonprofit (and sometimes subsidized) financial institutions. But, naturally, the causes of poverty may be multifaceted. Thus, uncoordinated programs across different ministries may fail to provide the right bundle of interventions that a household would need to improve their living standard. This lack of coordination in itself poses an interesting research agenda to understand the impacts of these multiapproach programs.

One recent example of such a program is the “graduation” approach—an integrated, multi-faceted program with livelihood promotion at its core that aims to “graduate” individuals out of extreme poverty and onto a long term, sustainable higher consumption path. BRAC, the world’s largest nongovernmental organization, has scaled-up this program in Bangladesh (Bandiera et al., 2016), while NGOs around the world have engaged in similar livelihood-based efforts. Six randomized trials across the world (Ethiopia, Ghana, Honduras, India, Pakistan, and Peru) found that the integrated multifaceted program was “sufficient” to increase long-term income, where long term is defined as 3 years after the productive asset transfer (Banerjee et al., 2015a,b,c). The results from the pooled analysis across all six countries found that the program led to sustainable and significant impacts in 10 out of 10 categories of impact. Using an index approach to account for multiple hypotheses testing, positive impacts were found for consumption, income and revenue, asset wealth, food security, financial inclusion, physical health, mental health, labor supply, political involvement, and women’s decision-making after 2 years. After a third year, the results remained the same in 8 out of 10 outcome categories (with point estimates falling to below statistical significance for physical health and women’s empowerment). Furthermore, the West Bengal site found even larger treatment effects after 7 years (Banerjee et al., 2016). The pattern of results are strikingly similar to the Bangladesh study (Bandiera et al., 2016).

These results are promising in that they show that a sufficient set of interventions is capable of alleviating poverty sustainably and are thus important for policy. They should whet the appetite, both for a more theoretically grounded understanding of exactly which market failures led to a poverty trap, as well as a more practically grounded understanding of whether all of the interventions were truly necessary or if certain components could be removed. In the event that some components are unnecessary, costs could be lowered considerably, allowing the program to reach more people using the same budget. Returning to the theme of this paper, there are two complementary methods to tackle testing the important mechanisms behind the theory, and success or failure, of these programs: data and experimental design.

The ideal method, if unconstrained by budget and organizational constraints, is a complex experimental design that randomizes all permutations of each component. The productive asset transfer, if the only issue were a credit market failure, may have been sufficient to generate these results, and if no other component enabled an individual to accumulate sufficient capital to acquire the asset, the transfer alone may have been a

necessary component. The savings component on the other hand may have been a substitute for the productive asset transfer, by lowering transaction costs to save and serving as a behavioral intervention, which facilitated staying on task to accumulate savings. Clearly, it is not realistic in one setting to test the necessity or sufficiency of each component and interaction across components: Even if treated simplistically with each component either present or not, this would imply  $2 \times 2 \times 2 \times 2 = 16$  experimental groups.

Data can also provide important insights, even absent experimental design variation. Take the savings component, for example. For the savings component to be either a necessary or sufficient component, presumably an increase in the flow of savings must be observed (but not necessarily the stock, since withdrawals for investment purposes may bring the stock back down). The evidence from the graduation programs shows widely varying impacts on savings, far more than the results of the program itself. For example, in the most extreme case, savings increased in Ethiopia by purchasing power parity (PPP) US\$707<sup>21</sup> compared to only PPP US\$17 in Ghana. This suggests that savings may be an important component but is neither a necessary nor sufficient component for some level of success.

Several studies have tackled pieces of the puzzle. The way forward is going to be the development of a mosaic of these studies that tests each component, but also includes sufficient contextual and market variations that it can help set policy for a myriad of countries and populations. For example, in a postconflict setting in Uganda, an NGO-led program provided youth groups with training and cash (US\$150) toward nonagricultural self-employment activity and found a 57% increase in business assets, 17% increase in work hours, and 38% increase in earnings 4 years after the cash grants (Blattman et al., 2014a). This program differs from the previously mentioned graduation programs in three potentially important and illuminating dimensions: postconflict versus nonpostconflict, youth versus general population of the extreme poor, group-level intervention versus household level, and no inclusion of ancillary components such as life coaching, savings, and health care. The first two differences speak to the applicability of the program to alternative sample frames and settings, whereas the third and fourth program variations suggest that either the driver of the impact of the program lies with the cash grants and training not the other components or that the group-level aspect improves the impact and effectively substitutes for the other components.

A second study in Uganda sheds insight into the value of the group-level intervention, as it randomly varies the group aspect of the intervention, as well as the intensity of supervision (Blattman et al., 2014b). These programs, as with the earlier Uganda program, differ from the graduation studies in that they do not include savings, health and

<sup>21</sup> Although note that the Ethiopia program design included a much stronger push for savings than the other programs, with savings put forward as almost a “mandatory” component, even though there was no consequence if households did not save.

life-coaching components, and are focused on enterprise development (rather than animal husbandry, the dominant livelihood in the graduation studies).

Thus, the initial studies discussed have established a base case that there exists a sufficient intervention package that increases long-term income. We highlight four lines of inquiry to understand more about the underlying mechanisms. First, long-term impacts are critical for assessing whether the short-run interventions actually addressed the underlying problems or rather just lasted a bit longer than a cash transfer. For example, graduation programs typically last 2 years while the graduation studies cited earlier measured impacts 3 years after the assets were transferred. If the household visits were a critical component in driving the observed impacts, longer-term measurement would be important to assess whether the behavioral changes motivated by the household visits persisted for more than just 1 year after the household visits ceased.

Second, as some of the previous studies have begun to do, more work is needed to tease apart the different components: asset transfer (addresses capital market failures), savings account (lowers savings transaction fee), information (addresses information failures), life-coaching (addresses behavioral constraints, and perhaps changes expectations and beliefs about possible return on investment), health services and information (addresses health market failures), consumption support (addresses nutrition-based poverty traps), etc. There will be no simple answer to the aforementioned queries, but further work can help isolate the conditions under which each of these components should be deemed necessary to address. And furthermore, for several of these questions, there are key open issues for *how* to address them; for example, life coaching can take on an infinite number of manifestations. Some organizations conduct life coaching through religion, others through interactive problem solving, and others through psychotherapy approaches ([Bolton et al., 2003, 2007; Patel et al., 2010](#)). Much remains to be learned not just about the promise of such life-coaching components, but how to make them work (if they work at all).

Third, general equilibrium effects should be considered, particularly as the programs are taken to scale. Here, the first task is to be more specific in data collection, as general equilibrium effects encompass a wide variety of indirect effects, such as price of transferred assets; spillovers from explicit sharing of granted resources; and increased economic activity from increasing the poor's wealth. A typical experimental design would either randomize across and within villages (assuming that the village is the boundary for generating general equilibrium effects) or for some issues, examining spillovers to nonparticipants in treatment versus control (as in [Angelucci and De Giorgi, 2009](#)).

Fourth, important lessons can be learned from understanding the consumption path taken by households after participating in these programs. The graduation program, for example, found important and cost-effective, but still modest, increases in long-term consumption. This finding suggests that households are not caught in an extreme poverty trap, where one simply needs to get households over a particular hump and they will

immediately converge to the equivalent of the middle class. Further work is needed to understand the long-term dynamics of such programs and what can be done to further increase income mobility.

## 5.2 Building long-term financial assets—pensions

Noncontributory pension programs are an important form of social protection in many developing countries—such as Brazil, South Africa, India, etc.,—and given the shift in demographics and the risk of poverty for the elderly ([UNDESA, 2013](#)), they are likely to grow in importance. The programs vary in shape and form: [Rofman et al. \(2015\)](#) compare pension programs across 14 different Latin American countries, showing differences in payment sizes, in timing of payment, in whether the pensions are targeted, etc. While to the best of our knowledge, there are few experimental studies of pension programs in developing countries.<sup>22</sup> RCTs can help us understand how differences in these design choices affect the labor market choices of working-age adults, retirement age, saving patterns, and how funds are used within the household.

## 6. IDEAS ONLY GO SO FAR: IMPLEMENTATION MATTERS TOO

A transfer program may look like a winner on paper but may be a total flop in practice if the implementation is haphazard. Some of this may be purely administrative, e.g., ensuring that the right number of staff is hired, and that they are properly trained and motivated. This may require incentives to not shirk on the job, as well as to not engage in bad behaviors, e.g., siphon off funds or food, or reallocate the funds to friends or political supporters rather than those who are most in need.

Therefore, in designing randomized evaluations of antipoverty programs, it is also important to think about whether, theoretically, a particular aspect of the implementation is likely to be particularly vulnerable to problems. There are two types of variations one can think about: (1) evaluations that vary the underlying structure of the program and (2) evaluations that layer on complementary actions that can be undertaken to improve program implementation given a fixed program design.

Experimentally varying the core elements of the underlying structure of a program is challenging, especially if aspects of the program have been written into law. However, the details of the underlying structure—from who should implement the program to how one should make the transfers—may matter tremendously, affecting the level of leakages and corruption, the targeting, the costs for beneficiaries to access the program, and potentially how beneficiaries spend their entitlements. For example, there is an extensive work, in general, exploring how officials’ incentives affect their work output,

<sup>22</sup> Although as of the time of this chapter, there are several exciting ongoing studies in Chile and India.

but to our knowledge less work on how the incentives provided to officials affects transfer program delivery. Similarly, there is a strain of research that shows that the type of person recruited may affect government efficiency (see, for example, recent empirical evidence from [Ashraf et al., 2014](#); [Hanna and Wang, 2014](#)), but less specifically on how changing who is selected to implement transfer programs affects the ultimate outcomes of households. For future research, one could vary the salary and incentive structure (amount and conditions) for current workers—and during the recruitment of new workers—to explore how it affects targeting and delivery.

One important question is whether governments should even directly implement these programs, or whether they should contract out delivery mechanisms. [Banerjee et al. \(2014\)](#) experimentally vary whether local officials distribute a government-run subsidized rice program or whether private citizens also bid for the right to run the program. They find that the bidding reduces the price—markup that citizens pay, without reducing quality. Follow-up work can include testing out different ways of contracting out, from changing who is eligible to bid to how the bidding process occurs to how new implementers are reevaluated. Moreover, the bidding process in that paper focuses on local government provision (i.e., at the village level). Future research could also help shed light on whether the procurement process should be done at that local level where individuals possess local information about how to get things done in that village, or should it be done at a higher level of government (e.g., district or province level) where one may also benefit from economies of scale.

A nice series of recent papers tests whether the nature of the delivery mechanism in itself affects outcomes. For example, [Aker et al. \(2011\)](#) experimentally test for the impact of providing cash versus mobile money in a short-run transfer program in Niger. An innovative feature of their experimental design was to also have a treatment group that simply got cash *and a cell phone* and to net out potential effects of having a cell phone more generally from mobile money. Mobile money not only reduced the nonprofit's distribution costs but it also reduced the households' costs to pick up their entitlement. This second feature of mobile money may be particularly important if we believe high transaction costs induce beneficiary households "leave money" on the table ([Currie and Gahvari, 2008](#)). Importantly, they also showed that spending patterns changed due to mobile money, hypothesizing that it also conferred greater privacy over one's finances. Further testing these mechanisms in the context of larger government programs to understand longer-run effects would be an important extension of this work: For example, would the ease and potential secrecy of payments attract richer people to apply for these types of transfer programs? In the long run, would local officials who may have previously siphoned off cash during disbursements find other ways to "tax" citizens who now receive cash directly via mobile money?

An ambitious project by [Muralidharan et al. \(2014\)](#) also aims to address some of these types of questions. They evaluated the impact of biometrically authenticated payments

infrastructure (“smart cards”) on beneficiaries of employment (NREGA) and pension (SSP) programs in Andhra Pradesh, India. The smart cards changed both how households collected their payments, as well as who was in charge of the cash distribution (as banks and technology service providers managed the new cash disbursal system). The state was rolling out the program across its 158 subdistricts, so the authors randomized which subdistricts were converted first. Following the introduction of the program, not only did the time it took beneficiaries to collect a payment fall, but also the delay in receiving the payment was reduced by almost 30%. The ease of payment induced households to actually work more. Households, thus, earned more, while payments to officials remained the same—hence, leakages fell quite dramatically. Similarly, an RCT conducted by [Banerjee et al. \(2014\)](#) shows that by simply asking local officials to input all of the names of the people who participated in NREGA into a database system to receive the funds transfer—i.e., increasing informational requirements for releasing funds and reducing the administrative tiers in the flow of funds process—led to a stark decline in leakages of public transfers, with no corresponding decrease in actual NREGA work.

Experimentally testing complementary programs that are layered on top of existing programs can also be important in improving the delivery of social protection programs. These programs do not necessarily require changing the existing program rules or functioning but instead provide additional information or services to help citizens better access their entitlements. For example, one could test how increased information on eligibility and program rules affects overall program leakages. [Ravallion et al. \(2013\)](#) do this: They experimentally vary whether beneficiaries see a half hour video on their entitlements under NREGA. They show that this form of information has very little impact on employment. Given that the form and level of information may matter, one may also test between varying types of information: for example, [Banerjee et al. \(2014\)](#) show that a card that informs households of their eligible status and entitlements reduce leakages in a subsidized rice program, and that making the card information public within the village has even larger impacts. Moreover, one can imagine experiments designed to test how providing households with direct help with their paperwork when applying affects who enrolls.

Questions about implementation at scale also relate to critical questions about the role of randomized trials given how they are often conducted in developing countries. For example, there is a broader debate about whether we would observe similar program results in NGO and government settings, given differences in implementation capacity between two (see, for example, [Bold et al., 2013; Dhaliwal and Hanna, 2014](#)). Of course, if one is evaluating a program with an NGO that will be scaled up by that NGO or similar ones, we may not particularly care if the program would look different if run by the government. However, often times we may also want to understand how evaluations with NGOs would differ in government and vice versa. Naturally, the fundamental problem here is of generalizability and sample size: a comparison of any one NGO to any one

government only compares that of NGO to that government. NGOs are not a monolithic set of institutions, and neither are governments. Thus it may be wrong to ask whether a government is better, or worse, at implementing than an NGO, and be more appropriate to ask whether “an” institution with certain specific characteristics or in certain specific cultural or political environments will be better at implementing than an institution with a different set of specific characteristics or environmental factors.

Treatment effects may depend on institutional type (government or NGO) for two broad reasons: behavioral responses and implementation efficiency. In terms of behavioral responses, treatment effects may depend upon the legitimacy of who delivers the program. In the specific case of social protection programs, how we expect households to respond to a particular transfer of food or cash is unlikely to change based on who is distributing it. But, how people respond to the specifics of the program may matter. For example, in the study in Morocco with labeled cash transfers ([Benhassine et al., 2015](#)), it could be that such labels only work from trusted and well-known institutions. Again, this is less of an issue over whether the NGO or government is delivering the service but about the overall level of legitimacy of the institution. Thus, one interesting design would be to see if the response to the nudges changes when households are randomized to receive more or less information on how legitimate the organization has been in implementing these programs in the past.

In terms of implementation efficiency, one can also imagine that different types of organizations may have different strengths and weaknesses in terms of the types of programs that they can deliver. Suppose, for example, that citizens would be indifferent between cash and an in-kind transfer if both were implemented perfectly. However, one type of organization has strength in reducing the leakages in the delivery of cash relative to a second organization, and the second organization is better at reducing leakages in the in-kind transfer. Thus, citizens would ultimately prefer different types of transfers from different types of organizations due to the relative differences in leakages. Again, this preference may be symptomatic of a difference between government and an NGO but speaks to larger differences in the relative implementation abilities of different organizations and how can one improve upon their weaknesses.

## 7. CONCLUSION: KEY AREAS FOR FURTHER WORK

Since the innovative and instrumental randomized evaluation of Progresa in Mexico, there has been a burst of important and exciting experiments in this area. This has greatly informed our understanding of what can “work” in trying to redistribute to the poor, as well as what can reduce both behavioral constraints and market failures.

So, then the question becomes, where should we focus our research efforts next? We highlight three areas for further work, aside from those discussed earlier: interactions of demand and supply, long-term effects, and general equilibrium effects.

## 7.1 Key areas for further work

### 7.1.1 *Interactions of demand and supply*

A vital question is how transfer programs work across different contexts. For example, [Galiani and McEwan \(2013\)](#) document that the effect of the Honduran PRAF CCT program was much larger in the two poorest strata, with the effect not being statistically significant in the three richer areas.

This question is similar to the one at the center of the debate over who implements (e.g., a nonprofit or government): If one wants to understand how a program will work in a specific context, we may not care whether the evaluation findings are portable to another context. But if we want to understand whether a program would have similar results in another area, or if the results will change with policy changes in the current area, it is important to understand how the theoretically important underlying features of an area impact outcomes.

More broadly, there are lots of unanswered questions about the interaction of transfer programs with existing conditions, particularly supply-side conditions: For example, How does school quality or health-care availability affect the adherence to CCT conditions? Do food or other in-kind transfers work better than cash in areas with more limited food supplies? Do transfers facilitate access to finance by reducing risk to lenders? And so forth.

To answer these kinds of questions, one would ideally not only vary the introduction of a transfer program but would also cross this with an experimental change in a supply-side feature. For example, to isolate how increased health-care availability affects the adherence to CCT health conditions, one would randomize areas to four treatments: a pure control, CCT only, an increase in nurses only, and CCT and an increase in nurses.

An extension of this would be to test the effectiveness of different types of transfer programs under different conditions: For example, we might think that a UCT may be more effective at redistributing to the poor than a CCT in areas where there is limited health availability, since the inability to adhere to the conditions may scare off or reduce payments to the poor. Thus, rather than having just a pure control, one may want to compare CCTs to UCTs across areas with and without the induced increases in nurse availability.

### 7.1.2 *Long-term effects*

There is a tension between trying to measure a program's long-run impacts versus scaling up a "working" program to the control group. However, long-run impacts are important to measure, especially if there are reasons to believe that a program's impacts may evolve differently as time goes on (and potentially have general equilibrium effects as we discuss later).

For example, while we know quite a bit about the short-run impacts of different targeting methods, we know less about their relative long-run effects. Using

quasiexperimental variation, [Camacho and Conover \(2011\)](#) show that Colombia's targeting system was manipulated over time, as local officials better learned the rules of the game. One can imagine experimentally varying different targeting methods across different locations, and then repeating the same method in each respective location during the recertification process, to determine whether the relative efficacy of different methods change as both households and officials learn the systems over time.

Similarly, there are many questions about the long-run impacts of the transfers themselves: What happens to households after the transfers are complete? For example, Did CCTs achieve the goal of changing the outcomes the next generation, i.e., Did the children who attended school for longer, or had improved test scores, ultimately do better in the labor market? In the case of the graduation studies discussed earlier, Bangladesh and West Bengal, India sites have followed households for 7 years, and found that the positive treatment effects increased from 3–7 years ([Banerjee et al., 2016](#)).

### **7.1.3 General equilibrium effects**

With relatively large sums being distributed, antipoverty programs may have broader effects than one initially expects, with these effects being potentially quite large. These effects can take various forms. Antipoverty programs can affect insurance and lending markets within villages ([Angelucci and de Giorgi, 2009](#)), natural resource demand ([Gertler et al., 2013; Hanna and Oliva, 2015](#)), and labor markets ([Muralidharan et al., 2016](#)). Some of the effects could be positive, while others can be negative. While we touched on the idea of general equilibrium effects earlier, it is important enough of a topic to warrant its own section here.

Importantly, the general equilibrium effects may differ by type of transfers. For example, [Cunha et al. \(2013\)](#) document different effects on prices of consumer goods when villages have been randomized to cash versus in-kind transfer programs, particularly in remote areas. The general equilibrium effects may also vary across contexts: For example, CCTs induce the positive peer effects on the schooling outcomes of ineligible children in Mexico's Progresa ([Bobonis and Finan, 2009; Laliv and Cattaneo, 2009](#)) but no effects on ineligible children in the Honduran PRAF ([Galiani and McEwan, 2013](#)). At the more extreme, [Barrera-Osorio et al. \(2011\)](#) find negative spillovers: Siblings (particularly sisters) of CCT recipients are less likely to attend school and more likely to drop out.

While there have been a few studies, including those discussed earlier, that have tried to capture broader effects, this is still an area where our understanding is relatively sparse and where there is a need for further research. However, given the multitude of types of general equilibrium effects that may be possible, this is a case where we are particularly worried about multiple hypothesis testing. Careful theory-detailing predictions, coupled with prespecified hypotheses, may be important in building a robust model that successfully predicts outcomes in new settings, thus properly guides policy-making.

Identifying spillover effects should be built into the research plan when plausible and viable. Three basic approaches have been employed: (1) through experimental design: randomizing the density of treatment within a geographic area (or within any unit within which one expects there to be spillovers or general equilibrium effects), (2) through data collection on ineligibles: this is strengthened when combined with the first, but even on its own can shed important insights ([Angelucci and de Giorgi, 2009](#)), and (3) through data collection on process changes: for example, collecting specific data on informal transfers, credit and savings could identify behaviors that indicate the presence of general equilibrium effects.

## 7.2 Final thoughts

Putting these elements together poses its own challenges. Naturally, there is no simple diagnostic that assesses which markets are missing for a society, to then provide an easy prescription for which of the earlier programs to implement. And, on the other end of the spectrum, it is simply not practical to implement at scale a program that assesses for each individual what constraints they face and then provides the exact program that targets their particular situation. The policy challenge lies in how to find the policy that balances the operational constraints of scale with the targeting constraints of both identifying the poorest and minimizing the false positives, i.e., policies targeting an issue that are not relevant for a household or community. National social protection strategies should think holistically: How do specific policies interact with each other as either complements or substitutes? What populations are critically missing from existing policies? and how well do existing policies improve long-term outcomes so as to reduce the eventual tax burden on society?

## REFERENCES

- Aker, J.C., 2014. Comparing Cash and Voucher Transfers in a Humanitarian Context: Evidence from the Democratic Republic of Congo.
- Aker, J.C., Boumnijel, R., McClelland, A., Tierney, N., 2011. Zap it to Me: The Short-Term Impacts of a Mobile Cash Transfer Program. Center for Global Development. Working Paper No. 268.
- Alatas, V., Banerjee, A., Hanna, R., Olken, B.A., Purnamasari, R., Wai-Poi, M., 2013. Does Elite Capture Matter? Local Elites and Targeted Welfare Programs in Indonesia. Working Paper 18798. National Bureau of Economic Research. <http://www.nber.org/papers/w18798>.
- Alatas, V., Banerjee, A., Hanna, R., Olken, B.A., Tobias, J., 2012. Targeting the poor: evidence from a field experiment in Indonesia. Am. Econ. Rev. 102 (4), 1206–1240. <http://dx.doi.org/10.1257/aer.102.4.1206>.
- Alatas, V., Banerjee, A., Hanna, R., Olken, B., Purnamasari, R., 2016. Self-targeting: evidence from a field experiment in Indonesia. J. Polit. Econ. 124 (2), 371–427.
- Alzúa, M.L., Cruces, G., Ripani, L., 2013. Welfare programs and labor supply in developing countries: experimental evidence from Latin America. J. Popul. Econ. 26 (4), 1255–1284.
- Angelucci, M., De Giorgi, G., 2009. Indirect effects of an aid program: how do cash transfers affect ineligibles' consumption? Am. Econ. Rev. 99 (1), 486–508. <http://dx.doi.org/10.1257/aer.99.1.486>.

- Angelucci, M., Karlan, D., Zinman, J., 2015. Microcredit impacts: evidence from a randomized microcredit program placement experiment by compartamos banco. *Am. Econ. J.* 7 (1), 151–182. <http://dx.doi.org/10.1257/app.20130537>.
- Ashraf, N., Bandiera, O., Lee, S.S., 2014. Do-gooders and Go-getters: Career Incentives, Selection, and Performance in Public Service Delivery. Suntory and Toyota International Centres for Economics and Related Disciplines, LSE.
- Ashraf, N., Karlan, D., Yin, W., 2006. Tying Odysseus to the mast: evidence from a commitment savings product in the Philippines. *Q. J. Econ.* 121 (2), 673–697.
- Attanasio, O., Augsburg, B., De Haas, R., Fitzsimons, E., Harmgart, H., 2015. The impacts of microfinance: evidence from joint-liability lending in Mongolia. *Am. Econ. J.* 7 (1), 90–122. <http://dx.doi.org/10.1257/app.20130489>.
- Augsburg, B., De Haas, R., Harmgart, H., Meghir, C., 2015. The impacts of microcredit: evidence from Bosnia and Herzegovina. *Am. Econ. J.* 7 (1), 183–203. <http://dx.doi.org/10.1257/app.20130272>.
- Baird, S., McIntosh, C., Özler, B., 2011. Cash or condition? Evidence from a cash transfer experiment. *Q. J. Econ.* 126 (4), 1709–1753. <http://dx.doi.org/10.1093/qje/qjr032>.
- Baker, J.L., Grosh, M.E., 1994. Poverty reduction through geographic targeting: how well does it work? *World Dev.* 22 (7), 983–995. [http://dx.doi.org/10.1016/0305-750X\(94\)90143-0](http://dx.doi.org/10.1016/0305-750X(94)90143-0).
- Bandiera, O., Burgess, R., Das, N., Gulesci, S., Rasul, I., Sulaiman, M., 2016. Labor Markets and Poverty in Village Economies. LSE Working Paper. <http://sticerd.lse.ac.uk/dps/eopp/eopp43.pdf>.
- Banerjee, A., Duflo, E., Chattopadhyay, R., Shapiro, J., 2016. Long Term Impact of a Livelihood Intervention: Evidence from West Bengal (Working Paper).
- Banerjee, A., Duflo, E., Glennerster, R., Kinnan, C., 2015a. The miracle of microfinance? Evidence from a randomized evaluation. *Am. Econ. J.* 7 (1), 22–53. <http://dx.doi.org/10.1257/app.20130533>.
- Banerjee, A., Hanna, R., Kyle, J., Olken, B., Sumarto, S., 2014. Information Is Power: Identification Cards and Food Subsidy Programs in Indonesia (Working Paper).
- Banerjee, A., Karlan, D., Zinman, J., 2015b. Six randomized evaluations of microcredit: introduction and further steps. *Am. Econ. J.* 7 (1), 1–21.
- Banerjee, A., Newman, A., 1993. Occupational choice and the process of development. *J. Political Econ.* 101, 274–298.
- Banerjee, A.V., Hanna, R., Kreindler, G., Olken, B.A., 2015c. Debunking the Stereotype of the Lazy Welfare Recipient: Evidence from Cash Transfer Programs Worldwide. [http://papers.ssrn.com/sol3/papers.cfm?abstract\\_id=2703447](http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2703447).
- Bardhan, P.K., Mookherjee, D., 2000. Capture and governance at local and national levels. *Am. Econ. Rev.* 90 (2), 135–139. <http://dx.doi.org/10.1257/aer.90.2.135>.
- Barrera-Osorio, F., Bertrand, M., Linden, L., Perez-Calle, F., 2011. Improving the design of conditional transfer programs: evidence from a randomized education experiment in Colombia. *Am. Econ. J.* 3 (2), 167–195.
- Barrera-Osorio, F., Filmer, D., 2013. Incentivizing Schooling for Learning: Evidence on the Impact of Alternative Targeting Approaches. World Bank Policy Research Working Paper, no. 6541.
- Beath, A., Christia, F., Enikolopov, R., 2013. Do elected councils improve governance? Experimental evidence on local institutions in Afghanistan. *MIT Political Sci. Dep. Res. Pap.* 2013 (24).
- Beaman, L., Karlan, D., Thuybaert, B., Udry, C., 2014. *Self-Selection into Credit Markets: Evidence from Agriculture in Mali* (Working Paper No. 20387). National Bureau of Economic Research. Retrieved from <http://www.nber.org/papers/w20387>.
- Benhassine, N., Devoto, F., Duflo, E., Dupas, P., Pouliquen, V., 2015. Turning a shove into a nudge? a ‘labeled cash transfer’ for education. *Am. Econ. J.* 7 (3), 86–125.
- Berry, J., 2014. Child Control in Education Decisions: An Evaluation of Targeted Incentives to Learn in India (Working Paper).
- Besley, T., Coate, S., 1991. Public provision of private goods and the redistribution of income. *Am. Econ. Rev.* 81 (4), 979–984.
- Besley, T., Coate, S., 1992. Welfare versus welfare: incentive arguments for work requirements in poverty-alleviation programs. *Am. Econ. Rev.* 82 (1), 249–261.

- Blattman, C., Fiala, N., Martinez, S., 2014a. Generating skilled self-employment in developing countries: experimental evidence from Uganda. *Q. J. Econ.* 129 (2), 697–752.
- Blattman, C., Green, E., Annan, J., Jamison, J., 2014b. The Returns to Cash and Microenterprise Support Among the Ultra-poor: A Field Experiment. Columbia University. Working Paper. [http://papers.ssrn.com/sol3/papers.cfm?abstract\\_id=2439488](http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2439488).
- Bobonis, G.J., Finan, F., 2009. Neighborhood peer effects in secondary school enrollment decisions. *Rev. Econ. Statistics* 91 (4), 695–716.
- Bold, T., Kimenyi, M., Mwabu, G., Ng'ang'a, A., Sandefur, J., 2013. Scaling Up What Works: Experimental Evidence on External Validity in Kenyan Education. Center for Global Development. Working Paper, no. 321.
- Bolton, P., Bass, J., Betancourt, T., Speelman, L., Onyango, G., Clougherty, K.F., Neugebauer, R., Murray, L., Verdeli, H., 2007. Interventions for depression symptoms among adolescent survivors of war and displacement in northern Uganda: a randomized controlled trial. *JAMA* 298 (5), 519. <http://dx.doi.org/10.1001/jama.298.5.519>.
- Bolton, P., Bass, J., Neugebauer, R., Verdeli, H., Clougherty, K.F., Wickramaratne, P., Speelman, L., Ndogoni, L., Weissman, M., 2003. Group interpersonal psychotherapy for depression in rural Uganda: a randomized controlled trial. *JAMA* 289 (23), 3117–3124.
- Bryan, G., Karlan, D., Nelson, S., 2010. Commitment devices. *Ann. Rev. Econ.* 2 (1), 671–698. <http://dx.doi.org/10.1146/annurev.economics.102308.124324>.
- Bursztyn, L., Coffman, L.C., 2012. The schooling decision: family preferences, intergenerational conflict, and moral hazard in the Brazilian favelas. *J. Political Econ.* 120 (3), 359–397. <http://dx.doi.org/10.1086/666746>.
- Camacho, A., Conover, E., 2011. Manipulation of social program eligibility. *Am. Econ. J. Econ. Policy* 3 (2), 41–65.
- Cameron, A.C., Miller, D.L., 2010. Robust Inference with Clustered Data. University of California, Department of Economics. Working Papers. <http://www.econstor.eu/handle/10419/58373>.
- Chen, S., Ravallion, M., Wang, Y., 2006. Di Bao: A Guaranteed Minimum Income in China's Cities?, Vol. 3805. World Bank Publications, Washington, D.C.
- Chetty, R., Looney, A., 2006. Consumption smoothing and the welfare consequences of social insurance in developing economies. *J. Public Econ.* 90 (12), 2351–2356. <http://dx.doi.org/10.1016/j.jpubeco.2006.07.002>.
- Christian, P., 2014. The Distributional Consequences of Group Procurement: Evidence from a Randomized Trial of a Food Security Program in Rural India (Working Paper).
- Coady, D., 2004. Designing and Evaluating Social Safety Nets: Theory, Evidence, and Policy Conclusions. Food Consumption and Nutrition. Division Discussion Paper No. 172.
- Cohen, J., Dupas, P., 2010. Free distribution or cost-sharing? evidence from a randomized malaria prevention experiment. *Q. J. Econ.* 125 (1), 1–45. <http://dx.doi.org/10.1162/qjec.2010.125.1.1>.
- Covarrubias, K., Davis, B., Winters, P., 2012. From protection to production: productive impacts of the Malawi social cash transfer Scheme. *J. Dev. Eff.* 4 (1), 50–77.
- Crépon, B., Devoto, F., Duflo, E., Pariente, W., 2015. Estimating the impact of microcredit on those who take it up: evidence from a randomized experiment in Morocco. *Am. Econ. J. Appl. Econ.* 7 (1), 123–150. <http://dx.doi.org/10.1257/app.20130535>.
- Cunha, J.M., De Giorgi, G., Jayachandran, S., 2013. The Price Effects of Cash Versus In-kind Transfers.
- Cunha, J.M., 2014. Testing Paternalism: Cash versus In-Kind Transfers. *American Economic Journal: Applied Economics* 6 (2), 195–230. <http://dx.doi.org/10.1257/app.6.2.195>.
- Currie, J., Gahvari, F., 2008. Transfers in cash and in-kind: theory meets the data. *J. Econ. Literature* 46 (2), 333–383.
- de Janvry, A., Fafchamps, M., Sadoulet, E., 1991. Peasant household behaviour with missing markets: some paradoxes explained. *Econ. J.* 101 (409), 1400. <http://dx.doi.org/10.2307/2234892>.
- de Mel, S., McKenzie, D., Woodruff, C., 2008. Returns to capital in microenterprises: evidence from a field experiment. *Q. J. Econ.* 123 (4), 1329–1372.

- Dhaliwal, I., Hanna, R., 2014. Deal with the Devil: The Successes and Limitations of Bureaucratic Reform in India. National Bureau of Economic Research. Working Paper 20482. <http://www.nber.org/papers/w20482>.
- Duflo, E., Dupas, P., Kremer, M., 2010. Education and Fertility: Experimental Evidence from Kenya (Working Paper).
- Duflo, E., Gale, W., Liebman, J., Orszag, P., Saez, E., 2006. Saving incentives for low- and middle-income families: evidence from a field experiment with H&R block. *Q. J. Econ.* 121 (4), 1311–1346.
- Dupas, P., Hoffmann, V., Kremer, M., Zwane, A.P., 2013. Micro-ordeals, Targeting, and Habit Formation (Working Paper).
- Edmonds, E., Schady, N., 2012. Poverty alleviation and child labor. *Am. Econ. J. Econ. Policy* 4 (4), 100–124.
- Elbers, C., Fujii, T., Lanjouw, P., Özler, B., Yin, W., 2007. Poverty alleviation through geographic targeting: how much does disaggregation help? *J. Dev. Econ.* 83 (1), 198–213.
- Epple, D., Romano, R., 2008. Educational vouchers and cream skimming. *Int. Econ. Rev.* 49 (4), 1395–1435.
- Fernald, L.C.H., Hidrobo, M., 2011. Effect of Ecuador's Cash Transfer Program (Bono de Desarrollo Humano) on Child Development in Infants and Toddlers: a Randomized Effectiveness Trial. *Soc. Sci. Med.* 72 (9), 1437–1446. <http://dx.doi.org/10.1016/j.socscimed.2011.03.005>.
- Ferre, C., 2009. Age at First Child: Does Education Delay Fertility Timing? the Case of Kenya. Social Science Research Network, Rochester, NY. SSRN Scholarly Paper ID 1344718. <http://papers.ssrn.com/abstract=1344718>.
- Fiszbein, A., Schady, N.R., 2009. Conditional Cash Transfers: Reducing Present and Future Poverty. World Bank, Washington, DC. A World Bank Policy Research Report 47603.
- Fudenberg, D., Levine, D., 2006. A dual self model of impulse control. *Am. Econ. Rev.* 96 (5), 1449–1476.
- Galiani, S., McEwan, P.J., 2013. The heterogeneous impact of conditional cash transfers. *J. Public Econ.* 103 (C), 85–96.
- Gentilini, U., Honorati, M., Yemtsov, R., 2014. The State of Social Safety Nets 2014. World Bank Group, Washinton, DC.
- Gertler, P.J., Martinez, S.W., Rubio-Codina, M., 2012. Investing cash transfers to raise long-term living standards. *Am. Econ. J. Appl. Econ.* 4 (1), 164–192.
- Gertler, P., Shelef, O., Wolfram, C., Fuchs, A., 2013. How Pro-poor Growth Affects the Demand for Energy. National Bureau of Economic Research. Working Paper 19092. <http://www.nber.org/papers/w19092>.
- Glewwe, P., Park, A., Zhao, M., 2014. A Better Vision for Development: Eyeglasses and Academic Performance in Rural Primary Schools in China. University of Minnesota Center for International Food and Agricultural Policy. Working Paper WP12-2.
- Grenier, J., Pattanayak, C.W., 2011. Randomized evaluation in legal assistance: what difference does representation (offer and actual use) make. *Yale LJ* 121, 2118.
- Hanna, R., Oliva, P., 2015. The effect of pollution on labor supply: evidence from a natural experiment in Mexico city. *J. Public Econ.* 122, 68–79.
- Hanna, R., Wang, S.-Y., 2014. Dishonesty and Selection into Public Service: Evidence from India.
- Hidrobo, M., Hoddinott, J., Peterman, A., Margolies, A., Moreira, V., March 2014. Cash, food, or vouchers? evidence from a randomized experiment in northern Ecuador. *J. Dev. Econ.* 107, 144–156. <http://dx.doi.org/10.1016/j.jdeveco.2013.11.009>.
- Hidrobo, M., Peterman, A., Heise, L., 2013. The Effect of Cash, Vouchers and Food Transfers on Intimate Partner Violence: Evidence from a Randomized Experiment in Northern Ecuador. International Food Policy Research Institute, Washington, DC. [https://www.wfp.org/sites/default/files/IPV-Hidrobo-Peterman\\_Heise\\_IPV%20Ecuador%203%2028%2014.pdf](https://www.wfp.org/sites/default/files/IPV-Hidrobo-Peterman_Heise_IPV%20Ecuador%203%2028%2014.pdf).
- Hoddinott, J., Sandström, S., Upton, J., 2014. The Impact of Cash and Food Transfers: Evidence from a Randomized Intervention in Niger. Available at: SSRN 2423772. [http://papers.ssrn.com/sol3/papers.cfm?abstract\\_id=2423772](http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2423772).
- Jacoby, H.G., 1997. Self-selection and the redistributive impact of in-kind transfers: an econometric analysis. *J. Hum. Resour.* 233–249.

- Jessee, C., Prencipe, L., Sherman, D., Banda, A., Ndiyoi, L., Tembo, N., Daidone, S., et al., 2013. Zambia's Child Grant Program: 24-Month Impact Report. American Institutes for Research.
- Karlan, D., Zinman, J., 2011. Microcredit in theory and practice: using randomized credit scoring for impact evaluation. *Science* 332 (6035), 1278–1284. <http://dx.doi.org/10.1126/science.1200138>.
- Kazianga, H., De Walque, D., Alderman, H., 2012. Educational and child labour impacts of two food-for-education schemes: evidence from a randomised trial in rural Burkina Faso. *J. Afr. Econ.* 21 (5), 723–760.
- Kling, J.R., Liebman, J.B., Katz, L.F., Sanbonmatsu, L., 2004. Moving to Opportunity and Tranquility: Neighborhood Effects on Adult Economic Self-sufficiency and Health from a Randomized Housing Voucher Experiment. [http://papers.ssrn.com/sol3/papers.cfm?abstract\\_id=588942](http://papers.ssrn.com/sol3/papers.cfm?abstract_id=588942).
- Kremer, M., Miguel, E., Thornton, R., 2009. Incentives to learn. *Rev. Econ. Statistics* 91 (3), 437–456.
- Kuziemko, I., Norton, M.I., Saez, E., Stantcheva, S., 2015. How elastic are preferences for redistribution? Evidence from randomized survey experiments. *Am. Econ. Rev.* 105 (4), 1478–1508. <http://dx.doi.org/10.1257/aer.20130360>.
- Lalive, R., Cattaneo, M.A., 2009. Social interactions and schooling decisions. *Rev. Econ. Statistics* 91 (3), 457–477.
- Leroy, J.L., Gadsden, P., González de Cossío, T., Gertler, P., 2013. Cash and in-kind transfers lead to excess weight gain in a population of women with a high prevalence of overweight in rural Mexico. *J. Nutr.* 143 (3), 378–383.
- Ma, X., Sylvia, S., Boswell, M., Rozelle, S., 2013. Ordeal Mechanisms and Information in the Promotion of Health Goods in Developing Countries: Evidence from Rural China. Rural Education Action Project Working Paper No. 266.
- McKenzie, D., Woodruff, C., 2006. Do entry costs provide an empirical basis for poverty traps? Evidence from Mexican microenterprises. *Econ. Dev. Cult. Change* 55 (1), 3–42. <http://dx.doi.org/10.1086/505725>.
- Meager, R., 2015. Understanding the Impact of Microcredit Expansions: A Bayesian Hierarchical Analysis of 7 Randomised Experiments. Available at: SSRN 2620834. [http://papers.ssrn.com/sol3/papers.cfm?abstract\\_id=2620834](http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2620834).
- Merttens, F., Hurrell, A., Marzi, M., Attah, R., Farhat, M., Kardan, A., MacAuslan, I., 2013. Kenya Hunger Safety Net Programme Monitoring and Evaluation Component. Impact Evaluation Final Report. Oxford Policy Management.
- Meyer, B.D., 1995. Lessons from the US unemployment insurance experiments. *J. Econ. Literature* 33 (1), 91–131.
- Mickelwright, J., Nagy, G., 2010. The effect of monitoring unemployment insurance recipients on unemployment duration: evidence from a field experiment. *Labour Econ.* 17 (1), 180–187.
- Miguel, E., Camerer, C., Casey, K., Cohen, J., Esterling, K.M., Gerber, A., Glennerster, R., et al., 2014. Promoting transparency in social science research. *Science* 343 (6166), 30–31.
- Muralidharan, K., Niehaus, P., Sukhtankar, S., 2014. Building State Capacity: Evidence from Biometric Smartcards in India. NBER. Working Paper No 19999.
- Muralidharan, Karthik, Paul Niehaus, Sandip Sukhtankar, 2016. General Equilibrium Effects of (Improving) Public Employment Programs: Experimental Evidence from India. Mimeo.
- Nichols, A.L., Zeckhauser, R.J., 1982. Targeting transfers through restrictions on recipients. *Am. Econ. Rev.* 372–377.
- Norton, M.I., Ariely, D., 2011. Building a better America—one wealth quintile at a time. *Perspect. Psychol. Sci.* 6 (1), 9–12. <http://dx.doi.org/10.1177/1745691610393524>.
- Olken, B., 2007. Monitoring corruption: evidence from a field experiment in Indonesia. *J. Political Econ.* 115, 200–249.
- Olken, B.A., 2015. Promises and perils of pre-analysis plans. *J. Econ. Perspect.* 29 (3), 61–80.
- Osili, Okonkwo, U., Long, B.T., 2008. Does female schooling reduce fertility? evidence from Nigeria. *J. Dev. Econ.* 87 (1), 57–75. <http://dx.doi.org/10.1016/j.jdeveco.2007.10.003>.
- Ozier, O., 2011. The Impact of Secondary Schooling in Kenya: A Regression Discontinuity Analysis. Working Paper. [http://economics.ozier.com/owen/papers/ozier\\_JMP\\_20110117.pdf](http://economics.ozier.com/owen/papers/ozier_JMP_20110117.pdf).

- Patel, V., Weiss, H.A., Chowdhary, N., Naik, S., Pednekar, S., Chatterjee, S., De Silva, M.J., Bhat, B., Araya, R., King, M., 2010. Effectiveness of an intervention led by lay health counsellors for depressive and anxiety disorders in primary care in Goa, India (MANAS): a cluster randomised controlled trial. *Lancet* 376 (9758), 2086–2095.
- Paxson, C., Schady, N., 2010. Does money matter? the effects of cash transfers on child development in rural Ecuador. *Econ. Dev. Cult. Change* 59 (1), 197–229.
- Ravallion, M., 1991. Reaching the rural poor through public employment: arguments, evidence, and lessons from South Asia. *World Bank Res. Observer* 6 (2), 153–175.
- Ravallion, M., Van de Walle, D.P., Dutta, P., Murgai, R., 2013. Testing Information Constraints on India's Largest Antipoverty Program. World Bank Policy Research. Working Paper, no. 6598. [http://papers.ssrn.com/sol3/papers.cfm?abstract\\_id=2323980](http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2323980).
- Rawls, J., 1971. *A Theory of Justice*, Rev ed. Belknap Press of Harvard University Press, Cambridge, Mass.
- Rofman, R., Apella, I., Vezza, E., 2015. Beyond Contributory Pensions: Fourteen Experiences with Coverage Expansion in Latin America. World Bank Publications. <https://books.google.com/books?hl=en&lr=&id=5GOzBQAAQBAJ&oi=fnd&pg=PP1&dq=Beyond+Contributory+Pensions:+Fourteen+Experiences+with+Cov>erage+Expansion+in+Latin+America&ots=OCI8ZK8PDK&sig=Bza3L1VoqX0fNtcDLvnI3KrOUY.
- Saavedra, J., Garcia, S., 2012. Impacts of Conditional Cash Transfer Programs on Educational Outcomes in Developing Countries. Product Page. [http://www.rand.org/pubs/working\\_papers/WR921-1.html](http://www.rand.org/pubs/working_papers/WR921-1.html).
- Schady, N., Araujo, M.C., Peña, X., López-Calva, L.F., 2008. Cash transfers, conditions, and school enrollment in Ecuador [with comments]. *Econ. J. Lat. Am. Caribb. Econ. Assoc.* 43–77.
- Schultz, T.P., 2004. School subsidies for the poor: evaluating the Mexican progresita poverty program. *J. Dev. Econ.* 74 (1), 199–250.
- Seabright, P., 1996. Accountability and decentralisation in government: an incomplete contracts model. *Eur. Econ. Rev.* 40 (1), 61–89.
- Singer, P., 1997. The drowning child and the expanding circle. In: New Internationalist, April, 28–30.
- Skoufias, E., Unar, M., de Cossio, T.G., 2013. The poverty impacts of cash and in-kind transfers: experimental evidence from rural Mexico. *J. Dev. Eff.* 5 (4), 401–429.
- Tarozzi, A., Desai, J., Johnson, K., 2015. The impacts of microcredit: evidence from Ethiopia. *Am. Econ. J. Appl. Econ.* 7 (1), 54–89. <http://dx.doi.org/10.1257/app.20130475>.
- The Kenya CT-OVC Evaluation Team, 2012. The impact of Kenya's cash transfer for Orphans and vulnerable children on human capital. *J. Dev. Eff.* 4 (1), 38–49.
- UNDESA, 2013. World Population Ageing Report. UN Department of Economic and Social Affairs.
- UNHCR, 2014. Statistical Online Population Database. United Nations High Commissioner for Refugees.
- Van den Berg, G.J., Van der Klaauw, B., 2006. Counseling and monitoring of unemployed workers: theory and evidence from a controlled social experiment. *Int. Econ. Rev.* 47 (3), 895–936.
- Vermeersch, C., Kremer, M., 2004. Schools Meals, Educational Achievement and School Competition: Evidence from a Randomized Evaluation. The World Bank.

## CHAPTER 8

# Social Experiments in the Labor Market

J. Rothstein<sup>\*§,a</sup>, T. von Wachter<sup>§¶,1</sup>

\*University of California, Berkeley, Berkeley, CA, United States

§NBER (National Bureau of Economic Research), Cambridge, MA, United States

¶University of California, Los Angeles, Los Angeles, CA, United States

<sup>1</sup>Corresponding author: E-mail: twachter@econ.ucla.edu

## Contents

1. Introduction	556
2. What Are Social Experiments? Historical and Econometric Background	560
2.1 A Primer on the history and topics of social experiments in the labor market	560
2.2 Social experiments as a tool for program evaluation	563
2.2.1 <i>The benchmark case: experiments with perfect compliance</i>	564
2.2.2 <i>Imperfect compliance and the local average treatment effect</i>	566
2.3 Limitations of the experimental paradigm	567
2.3.1 <i>Spillover effects and the stable unit treatment value assumption</i>	567
2.3.2 <i>Endogenously observed outcomes</i>	567
2.3.3 <i>Site and group effects</i>	568
2.3.4 <i>Treatment effect heterogeneity and external validity</i>	568
2.3.5 <i>Hidden treatments</i>	568
2.3.6 <i>Mechanisms and multiple treatments</i>	569
2.4 Quasiexperimental and structural research designs	569
3. A More Thorough Overview of Labor Market Social Experiments	570
3.1 Labor supply experiments	570
3.1.1 <i>The income maintenance experiments</i>	570
3.1.2 <i>Welfare reform experiments</i>	571
3.1.3 <i>Reemployment subsidy experiments</i>	577
3.2 Training experiments	578
3.3 Job search assistance	587
3.4 Practical aspects of implementing social experiments	594
4. Going Beyond Treatment–Control Comparisons to Resolve Additional Design Issues	596
4.1 Spillover effects and stable unit treatment value assumption	597
4.1.1 <i>Addressing the issue ex post</i>	597
4.1.2 <i>Addressing the issue ex ante through the design of the experiment</i>	599
4.2 Endogenously observed outcomes	600
4.2.1 <i>Addressing the issue ex post</i>	602
4.2.2 <i>Addressing the issue ex ante through the design of the experiment</i>	606

<sup>a</sup> We thank Ben Smith and Audrey Tiew for sterling research assistance, and Angus Deaton, Larry Katz, Jeff Smith, and conference participants for helpful comments.

<b>4.3</b>	Site and group effects	607
4.3.1	<i>Addressing the issue ex post</i>	608
4.3.2	<i>Addressing the issue ex ante through the design of the experiment</i>	612
<b>4.4</b>	Treatment effect heterogeneity and external validity	613
4.4.1	<i>Addressing the issue ex post</i>	613
4.4.2	<i>Addressing the issue ex ante through the design of the experiment</i>	616
<b>4.5</b>	Hidden treatments	618
4.5.1	<i>Addressing the issue ex post</i>	619
4.5.2	<i>Addressing the issue ex ante through the design of the experiment</i>	620
<b>4.6</b>	Mechanisms and multiple treatments	620
4.6.1	<i>Addressing the issue ex post</i>	621
4.6.2	<i>Addressing the issue ex ante through the design of the experiment</i>	627
<b>5.</b>	Conclusion	628
	References	630

## Abstract

Large-scale social experiments were pioneered in labor economics and are the basis for much of what we know about topics ranging from the effect of job training to incentives for job search to labor supply responses to taxation. Random assignment has provided a powerful solution to selection problems that bedevil nonexperimental research. Nevertheless, many important questions about these topics require going beyond random assignment. This applies to questions pertaining to both internal and external validity and includes effects on endogenously observed outcomes, such as wages and hours; spillover effects; site effects; heterogeneity in treatment effects; multiple and hidden treatments; and the mechanisms producing treatment effects. In this chapter, we review the value and limitations of randomized social experiments in the labor market, with an emphasis on these design issues and approaches to addressing them. These approaches expand the range of questions that can be answered using experiments by combining experimental variation with econometric or theoretical assumptions. We also discuss efforts to build the means of answering these types of questions into the *ex ante* design of experiments. Our discussion yields an overview of the expanding toolkit available to experimental researchers.

## Keywords

Labor market; Labor supply; Program evaluation; Social experiment

## JEL Codes

H53; I38; J22; J24; J31; J65

## 1. INTRODUCTION

There is a very long history of social experimentation in labor markets. Experiments have addressed core labor market topics, such as labor supply, job search, and human capital accumulation, and have been central to the academic literature and policy discussion, particularly in the United States, for many decades.

By many accounts, the first large-scale social experiment was the New Jersey Income Maintenance Experiment (IME), initiated in 1968 by the US Office of Economic

Opportunity (OEO) to test the effect of income transfers and income tax rates on labor supply. While many subsequent experiments have been designed to evaluate a single program or treatment, the IME was intended instead to map out a response surface. Participants were assigned to a control group or to one of eight treatment arms that varied in the income guarantee to a family that did not work and the rate at which this was taxed away as earnings rose. Three follow-up experiments—in rural North Carolina and Iowa; in Gary, Indiana; and in Seattle and Denver—with varying benefit levels and tax rates (and, in Seattle and Denver, a cross-cutting set of counseling and training treatments) were begun before data collection for the New Jersey experiment was complete.

Other early labor market experiments examined the effects of job search encouragement for Unemployment Insurance recipients; job training and job search programs; subsidized jobs for the hard-to-employ; and programs designed to push welfare recipients into work (Greenberg and Robins, 1986; Gueron, 2017, this volume). These topics have been explored repeatedly over the years since as researchers have sought to test new program designs or to build on the limitations of earlier research. There have also been many smaller-scale experiments, on bonus pay schemes, management structure, and other firm-level policies.<sup>1</sup>

From the beginning, the use of random assignment experiments (also known as randomized controlled trials, or RCTs) has been controversial in labor economics.<sup>2</sup> The primary, powerful appeal of RCTs is that they solve the assignment, or selection, problem in program evaluation. In nonexperimental studies (also known as “observational” studies), program participants may differ in observed and unobserved ways from those who do not participate, and econometric adjustments for this selection rely on unverifiable, often implausible assumptions (LaLonde, 1986; Fraker and Maynard, 1987; though see also Heckman and Hotz, 1989). With a well-executed randomization study, however, the treatment and control groups are comparable by design, making it straightforward to identify the effect of the treatment under study.

But set against this very important advantage are a number of drawbacks to experimentation. Early on, it was recognized that RCTs can be very expensive and hard to implement successfully. For example, it is not always possible to ensure that everyone assigned to receive a treatment receives a full dose, while those assigned to the control group receive none, although this is the experimental ideal. Sometimes it is not feasible to control participants’ behavior, and many participants deviate from their intended treatment assignments. In other cases, ethical, political, or operational considerations make

<sup>1</sup> We omit here audit studies aimed at uncovering discrimination in the labor market and elsewhere (e.g., Bertrand and Mullainathan, 2004; Kroft et al., 2013; Farber et al., 2015). These are covered by Bertrand and Duflo, elsewhere in this volume.

<sup>2</sup> For recent criticisms of reliance on RCTs with particular relevance to labor market studies, see Deaton (2010) and Heckman (2010). See also Heckman and Smith (1995).

it undesirable to limit access to alternative treatments. Although this can be partly addressed within the basic experimental paradigm, it does limit what can be learned.

More generally, while random assignment solves the assignment problem, it alone is not sufficient to resolve other problems that researchers often face. Many questions of interest can be answered only with something more than the familiar two-armed randomized control trial—a more complex experimental design, the augmentation of experimental data with additional, nonexperimental data, theoretically grounded assumptions, or a combination of these. We consider a number of such questions in this chapter. These include the following:

- *Questions about impacts on endogenously observed outcomes.* Consider the effect of job training on wages. Because wages are observed only for those who have jobs, and because training may affect the likelihood of working, the contrast in mean wages between randomly assigned treatment and control groups does not compare like to like and thus does not solve the assignment problem for this outcome.
- *Questions about spillovers and market-level impacts.* When one individual's outcome depends on others' treatment assignments, experimental estimates of treatment effects can be misleading about a program's overall effect. In the context of labor market programs, an increase in job search effort by a treatment group may lower the control group's job-finding chances, leading to an overstatement of the program's total effect (which will itself depend importantly on the scale at which the program is implemented). Similar issues can arise if subjects communicate with each other, leading to a dilution in treatment contrasts when access to information is part of the treatment.
- *Questions about heterogeneity of treatment effects.* Experiments have limited ability to identify heterogeneity of treatment effects, especially if heterogeneity is not fully characterized by well-defined observable characteristics. This is often of first-order importance, as in many cases the relevant question is not *whether* to offer a program (e.g., job training) but *for whom* to make it available, or *which* versions of the program are most effective (and why).
- *Questions about generalizability.* While in ideal cases experiments have high internal validity for the effect of the specific program under study on the specific experimental population, in the setting in which it is studied, they may have limited external validity for generalizations to other locations, to other programs (or even to other implementations of the same program), or to other populations. For example, a reemployment bonus program may have a very different effect in a full-employment local economy than when the local area is in a recession, or the same program offered in different sites may have dramatically different effects due to variation in local program administration or context.
- *Questions about mechanisms.* Many questions of interest in labor market research do not reduce to the effects of specific "treatments" on observed outcomes, but relate, at

best, to the mechanisms by which those effects arise. For example, an important question for the analysis of unemployment insurance (UI) programs is whether the unemployed are liquidity constrained or whether they can borrow or save to smooth consumption optimally across periods of employment and unemployment. And important questions about the design of welfare and disability policy turn on whether observed nonemployment is due to high disutility of work or to moral hazard. In each case, we want to distinguish income and substitution effects, a distinction that is in general not identified from the simple effect of a treatment on an observed outcome. Carefully designed experiments can shed light on the phenomena of interest, but may not be able to answer them directly.

To be clear, all of these questions are thorny under any methodological approach and are generally no easier to answer in quasiexperimental studies than in randomized experiments. One vocal group of critics of experimentation points to the importance of identifying the “structural” parameters—a full characterization of program enrollment decisions and the behavioral processes that lead to the observed outcomes—that determine program selection and impacts (see, e.g., Keane, 2010). In principle, many of the design issues above could indeed be avoided or addressed with estimates of the underlying structural parameters. But these structural parameters are difficult to measure. So-called structural methods generally trade off internal validity in pursuit of more external validity, but a study that fails to solve the assignment problem is unlikely to be any more generalizable than it is internally valid.

Unfortunately, while experiments can sometimes be designed to identify a few key structural parameters, or at least important combinations of them, it is rarely possible to design an experiment that directly identifies all of the structural parameters of interest. Thus, there can be value in combining the two paradigms. This involves imposing untestable assumptions about the processes of interest, while still resting on experimentation (or other empirical methods that offer high internal validity) where possible. The additional assumptions can dramatically enhance external validity if they are correct, though if they are incorrect—and this is generally untestable—both internal and external validity suffer.

The current frontier for labor market research—as in other fields—thus involves combining the best features of the two approaches to permit answers to more questions than are addressed by simple experiments, while retaining at least some of the credibility that these experiments can provide.

In this chapter, we discuss a variety of questions common in labor market research that require this sort of approach. We distinguish two broad strategies for answering these questions using experimental data. First, one can augment traditional randomized experiments by imposing additional structure, either economic or econometric, after the fact. In many cases, the amount of structure required, and the strength of the additional assumptions that are necessary, is small relative to the value of the results that can be

obtained. Our review gives a snapshot of an expanding toolkit with which researchers can address a wider range of questions based on variation from RCTs.<sup>3</sup>

The second broad strategy is to address the limitations of traditional experiments *ex ante*, via design of the experimental intervention or evaluation itself. In many cases, clever design choices—multiple treatment arms, carefully designed stratification, or randomization both across and within groups, for example—can allow for richer conclusions than would be possible via traditional experiments. This sort of approach has a long history—indeed, the very first large-scale social experiments, the IMEs of the late 1960s and early 1970s, can be seen as a version of it. But the pendulum swung away for a long time, and researchers have only recently begun to return to experimental designs that synthesize random experimental variation with more structural modeling. Recent examples of this approach include [Kling et al. \(2007\)](#), who use it to address potential biases from endogenous attrition, and [Crépon et al. \(2013\)](#), who quantify the importance of spillovers. In our view, approaches such as these represent the current research frontier.

The rest of this chapter proceeds as follows. In [Section 2](#), we give brief overviews of the history of social experiments in the labor market and of the value of RCTs for solving selection problems and summarize potential design issues that remain even with random assignment. In [Section 3](#), we review the types of programs and questions that have been analyzed, their main findings, and practical challenges that labor market experiments often confront. [Section 4](#) discusses approaches to addressing the design challenges from [Section 2](#) and thereby expanding the range of questions that can be answered. We discuss both *ex ante* and *ex post* approaches to resolving (or at least ameliorating) the issues. [Section 5](#) offers some concluding comments.

## **2. WHAT ARE SOCIAL EXPERIMENTS? HISTORICAL AND ECONOMETRIC BACKGROUND**

### **2.1 A Primer on the history and topics of social experiments in the labor market**

As the so-called “credibility revolution” has swept over empirical economics in the last generation, the role and status of experimental evidence have grown. Over the same period, the field of experimental economics has segmented—[List and Rasul \(2011\)](#) and [Harrison and List \(2004\)](#), for example, draw careful distinctions between social experiments and artificial, natural, and framed field experiments. Briefly, social experiments tend to be conducted at a large scale and to focus on the overall evaluation of

<sup>3</sup> This includes analyses of issues such as endogenously observed outcomes (e.g., [Ahn and Powell, 1993](#); [Groger, 2005](#); [Lee, 2009](#)); hidden treatments (e.g., [Kline and Walters, 2014](#); [Feller et al., 2014](#); [Pinto, 2015](#)); heterogeneous treatment effects (e.g., [Kline and Walters, 2014](#); [Heckman and Vytlacil, 2005](#)); and multiple treatments and mechanisms (e.g., [Card and Hyslop, 2005](#); [Schmieder et al., 2016](#); [DellaVigna et al., 2016](#)).

policies or programs, often already in place. By contrast, the various types of field experiments are typically smaller in scale and are more likely to use artificial treatments (e.g., behavioral games) that would not correspond directly to any specific policy but are designed primarily to uncover particular behavioral tendencies or parameters.

Although all of the many varieties of experiments have been used to study topics related to the labor market, this chapter focuses on large-scale social experiments, which in our view have had the largest impact on policy.

The social experiment/field experiment distinction corresponds roughly to the distinction drawn above between program evaluation and the identification of structural parameters—social experiments are, at root, evaluations of programs or policies, where field experiments are designed primarily to uncover one or more specific structural parameters.<sup>4</sup> As we discussed above, this distinction is less clear than it once was—scholars are increasingly drawing on program evaluation samples to understand structural relationships and using structural parameters to inform the design and interpretation of program evaluations. But while the distinction has been blurred, it has not been obliterated, and nearly all of the social experiments that we discuss in this chapter are designed, at least in part, to evaluate programs that either have been or might plausibly be implemented in roughly the form used in the experiment.

Another, related distinction has to do with the communities that conduct the different types of experiments. Social experiments are typically conducted at a large scale by an organization that specializes in this—historically, the “Big Three” players ([Greenberg and Shrader, 2004](#)) have been Mathematica, the Manpower Demonstration Research Corporation (MDRC), and Abt Associates—and has been hired by a government agency (most notably OPDR, the Office for Policy Development and Research within the Department of Labor’s Employment and Training Administration, and ASPE, the Assistant Secretary for Policy and Evaluation within the Department of Health and Human Services) or a large foundation (e.g., the Ford Foundation) for a specific study. By contrast, field experiments are more often overseen by individual scholars and their students, perhaps with the cooperation of a company or government agency that is not otherwise closely involved in the design.

The differences in the composition and organizational structure of social experimental and field experimental research teams relate to the scope of the work being carried out. A research team implementing a social experiment faces a number of practical and implementation challenges that are largely absent from laboratory experiments and closely related types of field experiments. Researchers rarely have access to a sampling frame corresponding to the population of interest; face practical, ethical, and political

<sup>4</sup> [Kling et al. \(2017\)](#), this volume, refer to experiments aimed at understanding mechanisms rather than at evaluating programs as “mechanism experiments.” [Gueron \(2017\)](#), this volume, discusses the tension between program evaluation and understanding mechanisms in early social experiments.

difficulties in randomly assigning access to treatment; have limited or no control over the specific implementation of the treatment, which is often under the control of an agency rather than the experimenter, or over the treatment alternatives that control participants may obtain and lack ready access to outcome measures for use in assessing the program's impact (or even to a well-defined set of outcomes of interest). Addressing these challenges often requires a large staff to collect pre- and posttreatment data, to minimize attrition between survey waves, and to monitor both the randomization of treatment and the fidelity of treatment delivery to the program model. The required scale is often out of the reach of individual researchers.

Most authors agree that the first large-scale social experiment in the labor market was the New Jersey IME (this is also known as the New Jersey Negative Income Tax Experiment), first initiated in 1968 and extended in various ways in other locations over the next several years. Consistent with the above dichotomy, this was a large-scale experiment that was initiated by the OEO, then an independent agency within the federal government that played a lead role in the War on Poverty. But in other ways it more closely resembles what would now be called a field experiment, albeit at a massive scale: it was first conceptualized by an individual researcher, Heather Ross, who proposed it to OEO, and it was designed not to evaluate a specific, well-developed program but to map out the surface of labor supply responses to a range of tax parameters and thereby to uncover semistructural economic parameters, the income and substitution effects of changes in tax rates.

Nearly all analyses of IME data went beyond simple treatment–control contrasts, using the data to estimate parametric or semiparametric labor supply models.<sup>5</sup> These models often incorporated corrections for the selection introduced by nonparticipation that relied on strong functional form assumptions (e.g., Tobits) and in some cases also rested on structural specifications of the response to nonlinear tax schedules. In many of these studies, the treatment and control groups were effectively pooled and it can be difficult to identify the extent to which the parameters are identified from experimental versus nonexperimental variation.

Another sense in which the IME diverged from much modern social experimental practice was in the source of outcome measures. The main outcome measures for the IME analyses were payments under the IME and labor supply measures drawn from participants' self-reports as part of the program's administration. But as in other experiments, many subjects failed to complete the follow-up surveys. Unfortunately, the design of the IME program meant that the private returns to continued reporting varied dramatically with both treatment status and

<sup>5</sup> Indeed, in 1990—7 years after the final experimental report from the follow-up Seattle–Denver Income Maintenance Experiment, and after many published analyses of the data—[Ashenfelter and Plant \(1990\)](#) are apparently the first to report the results as simple means by randomly assigned treatment group.

endogenous outcomes, as the income maintenance payments were made on the basis of these reports. Differential attrition made the results quite difficult to interpret ([Ashenfelter and Plant, 1990](#)).

In the wake of the IMEs, the field exploded. Greenberg et al. (1999; see also [Greenberg and Shroder, 2004](#)) identified 21 social experiments between 1962 and 1974, largely in education and health. By contrast, they identify 52 between 1975 and 1982 and 70 between 1983 and 1996, and most of these are directly related to the labor market. (There has not been as systematic a census of post-1996 experiments, but the pace of large-scale labor market experiments seems to have dropped off since then, at least in the United States. There has been rapid growth of social experiments in education over this period, however.) Greenberg et al. (1999; hereafter GSO) highlight important changes in the post-1975 experiments. In contrast to the IME, most involved only one or two treatment arms plus a control and were designed more as “black box” evaluations of the programs encapsulated in the treatments—often modifications on existing programs ([Gueron, 2017](#), this volume)—than as efforts to map out a response surface.

GSO emphasize that the vast majority of the experiments they identified focused on low-income populations, a fact that does not seem to have changed since their survey. Several topics stand out as central:

- *Human capital development.* Over one-third of the studies in GSO’s sample include at least one treatment arm involving a supported work experience, on-the-job training, vocational education or training, or basic education program, including General Educational Development (GED) programs.
- *Labor supply.* A number of experiments have involved interventions aimed at increasing labor supply, including the IMEs, studies of reemployment bonuses for UI recipients, and a broad group of welfare-to-work experiments conducted as part of the mid-1990s welfare reform movement.
- *Job search assistance.* Another common category of experiments examines interventions aimed at making disadvantaged workers’ job search efforts more effective, through counseling, job clubs, or job placement services.

These are not mutually exclusive. In particular, a number of programs and experiments combined job search assistance (JSA) with either job training or incentives to find work.

## 2.2 Social experiments as a tool for program evaluation

Random assignment solves the selection problem that often plagues nonexperimental program evaluations and makes it possible to generate uniquely credible evidence on the effects of well-defined, successfully implemented programs. In the absence of random assignment, people who participate in a program (those who are “treated”) are likely to differ in observed and unobserved ways from those who do not participate,

and the effect of this selection can be distinguished from the causal effect of the program only via the imposition of unverifiable assumptions about the selection process. This is a very important advantage of the experimental paradigm over other research methodologies (so-called “observational” comparisons), and we do not intend to minimize its contributions to the field of economics, public policy, and beyond.

But experiments have limitations as well—while they can have very high internal validity, at closer inspection this is true only for certain types of programs and certain types of outcomes; and even then there can be other challenges, such as difficulties in generalizing from the experimental results to a broader setting.

In this subsection, we discuss the value of experiments as a means of solving the selection problem. We then discuss some of the limitations of the experimental paradigm for program evaluation and policy analysis. Our discussion draws heavily on the [Angrist-Imbens-Rubin \(1996\)](#) “potential outcomes” framework. Some of the limitations we discuss can be addressed via careful design of the experimental study, while others require augmenting experimental methods with other tools. We take up these topics in [Section 4](#).

### **2.2.1 The benchmark case: experiments with perfect compliance**

The appeal of randomized experiments is that they make transparent the assumptions that permit causal inference and create a direct link between the implementation of the experiment and the key selection assumption. The simple contrast between those randomly assigned to participate in the program and those randomly excluded identifies the effect of being assigned to participate, subject only to the assumption that the randomization was conducted correctly. Moreover, in many cases this effect is identical to the effect of the program on its participants (known as the “effect of the treatment on the treated,” or TOT), which is often the main parameter of interest; in other cases, it is straightforward to convert the effect of assignment to participate (often known as the “intention to treat,” or ITT effect) into an estimate of the program treatment effect for a subpopulation of interest.

These results are well known (see, e.g., [Athey and Imbens, 2017](#), this volume), and we do not review them at length here. But it will be useful to have notation later. We use Donald Rubin’s potential outcomes framework for causal inference as set forth in [Holland \(1986\)](#). We consider the evaluation of a simple, well-defined program, such as an in-class job training course or a bonus scheme to encourage rapid return to work after a job displacement, where it is possible to assign individuals separately to participate or to be excluded from participation in the program.<sup>6</sup> For each individual  $i$ , one can imagine two possible outcomes: one that would obtain if  $i$  participated in the program,

<sup>6</sup> In the case of the bonus scheme, the “treatment” is eligibility for the bonus, not actual receipt.

$y_{i1}$ , and one that would obtain if he or she did not participate,  $y_{i0}$ .<sup>7</sup> The program's causal effect on person  $i$  is simply the difference between the outcome which would obtain if he/she participated and that which would obtain if she did not,  $\tau_i = y_{i1} - y_{i0}$ . When  $\tau_i > 0$ ,  $i$  would have a higher outcome if he/she participated than if he/she did not; when  $\tau_i < 0$ , the opposite is true.

Let  $D_i$  be an indicator for participation, with  $D_i = 1$  if  $i$  actually participates in the program and  $D_i = 0$  if  $i$  does not. The simplest estimator of the program's effect is the contrast between the average outcomes of those who participate and those who do not. This can be written as:

$$\begin{aligned} E[y_i | D_i = 1] - E[y_i | D_i = 0] &= E[y_{i1} | D_i = 1] - E[y_{i0} | D_i = 0] \\ &= E[\tau_i | D_i = 1] + (E[y_{i0} | D_i = 1] - E[y_{i0} | D_i = 0]). \end{aligned}$$

Thus, the simple participant–nonparticipant contrast combines two distinct components: the effect of the TOT,  $\tau^{\text{TOT}} = E[\tau_i | D_i = 1]$ , and a selection term,  $E[y_{i0} | D_i = 1] - E[y_{i0} | D_i = 0]$ , that captures the difference in outcomes that would have been observed between those who participated in the program and those who did not, had neither group participated (for example, had the program not existed). This second term arises because the process by which people select (or are selected) into program participation may generate differences between participants and nonparticipants other than their participation statuses. If so, the treatment–control difference cannot be interpreted as an estimate of the effect of the program.

In a simple social experiment,  $D_i$  is randomly assigned. This ensures that the distributions of  $y_{i0}$  and  $\tau_i$  are each the same for those with  $D_i = 0$  as for those with  $D_i = 1$ . The first implies that the selection term is zero; the second that the TOT effect equals the average treatment effect (ATE),  $E[\tau_i]$ , in the population represented by the study sample. Thus, the average causal effect is identified, not just in the treated subgroup but also in the larger population.<sup>8</sup>

This, in a nutshell, is the value of randomization in program evaluation. In a simple randomized control trial, the identification assumption that justifies causal inference is simply that the randomization was correctly executed. Of course, in any finite sample there may be differences in the sample averages of  $y_{0i}$  or  $\tau_i$  between treatment and control groups. But this variation is captured by the standard error of the experimental estimate. The estimate is unbiased, with measurable uncertainty, so long as the groups are the same in expectation.

<sup>7</sup> This notation rests on an assumption about the mechanisms by which the program operates, known as the “stable unit treatment value assumption,” or “SUTVA.” We discuss SUTVA at greater length below.

<sup>8</sup> This holds if the entire population of interest is part of the experiment. If the study sample is not representative of the broader population, the ATE identified will be local to the subpopulation represented by the sample.

## 2.2.2 Imperfect compliance and the local average treatment effect

A complication that often arises, and that will be central to some of our discussion below, is that it is not always possible to control subjects' program participation. Some subjects who are assigned to receive job training may not show up to their course, while others who are assigned to the control group, and thus not to receive training, may find another way into the program. This can be formalized by introducing an additional variable,  $Z_i$ , representing the experimenter's intention for individual  $i$ : an individual with  $Z_i = 1$  is intended to be served, and one with  $Z_i = 0$  is not to be.  $Z_i$  is related to  $D_i$ , but imperfectly: some (nonrandomly selected) individuals who are assigned  $Z_i = 1$  will wind up with  $D_i = 0$  (e.g., those who fail to arrive for their assigned training course), and others who are assigned  $Z_i = 0$  will wind up with  $D_i = 1$  (e.g., those who talk their way past the program screener).

With partial compliance, the experiment identifies neither the ATE nor the average effect of the TOT. Rather, the best that can be identified is the *local* average treatment effect, or LATE, for the subgroup of experimental subjects who comply with their experimental assignment. Specifically, let  $D_{i0}$  represent the individual's treatment status if assigned  $Z_i = 0$  and  $D_{i1}$  represent the treatment status if assigned  $Z_i = 1$ . The "complier" subpopulation is defined as those with  $D_{i0} = 0$  and  $D_{i1} = 1$ —those who receive the treatment if and only if they are assigned to receive it. The contrast between the average outcomes of those assigned to receive and not to receive treatment is then:<sup>9</sup>

$$E[y_i | Z_i = 1] - E[y_i | Z_i = 0] = \Pr\{D_{i0} = 0, D_{i1} = 1\} * E[\tau_i | D_{i0} = 0, D_{i1} = 1].$$

This is known as the ITT effect. The first term is the complier share of the experimental population; the second is the LATE for compliers.

In many cases, the ITT is the effect of primary interest. It represents the actual effect of offering access to the program in the setting in which the experiment takes place. Often, it is only possible to manipulate the option to participate (consider, for example, the offer of job training—one can never force individuals to participate in a training program), so the effect of manipulating this offer is the key parameter for evaluation of the programs under consideration.

In other cases, however, one might want to identify the effect of program participation (as distinct from the offer to participate). One can recover the LATE for compliers by dividing the ITT by the complier share, which can be identified as  $E[D_i | Z_i = 1] - E[D_i | Z_i = 0]$ ; equivalently, the LATE can be recovered from an instrumental variables regression using  $Z_i$  as an instrument for  $D_i$ .

<sup>9</sup> We assume here, as in nearly all analyses of experiments with partial compliance, that there are no "defiers" who receive the treatment if and only if they are assigned *not* to receive it ( $D_{i0} = 1$  and  $D_{i1} = 0$ ).

The LATE may differ from the ATE or even from the TOT. For example, in many settings, one would expect that people who receive the largest benefits from treatment will make disproportionate efforts to obtain it, even if assigned to the control group; in this case, the TOT will exceed the LATE. Unfortunately, the compliers are not always the population of primary interest. Further structure, or successful randomization of  $D_i$  itself, is required to identify the ATE or TOT.

## 2.3 Limitations of the experimental paradigm

The basic experimental paradigm is invaluable for its ability to resolve the fundamental problem of causal inference, by ensuring that estimated program effects are not confounded by selection into treatment. But it cannot solve all identification problems faced by program evaluators, nor answer all questions posed by labor economists seeking to understand the workings of the labor market. In the remainder of this section, we will briefly introduce six (partially overlapping) design issues that commonly arise in labor market experiments. In each case, identifying the effects of interest may require moving beyond the treatment–control contrast in outcomes from a simple randomized experiment. We discuss each in more detail in [Section 4](#), where we also discuss potential solutions to each.

### 2.3.1 Spillover effects and the stable unit treatment value assumption

The above brief overview of the econometrics of experiments glosses over an important assumption, known as the “stable unit treatment value assumption,” or SUTVA ([Angrist et al., 1996](#); [Athey and Imbens, 2017](#), this volume). Intuitively, this assumption states that the outcome of individual  $i$  is unaffected by the treatment status of each of the other study participants. Without this assumption, each individual has not 2 but  $2^N$  potential outcomes, making analysis intractable. For many program evaluations, SUTVA is innocuous. But in other cases it can be quite restrictive. For example, the provision of JSA to some individuals may create “congestion” in the labor market, reducing the job-finding rates of others participating in that market. This is a violation of SUTVA and will lead a simple randomized trial to overstate the total effect of JSA. Another potential violation of SUTVA occurs if members of the treatment group interact with each other or with the control group in a way that dilutes the treatment difference between them—for example, if the treatment involves information provision but treated individuals pass that information onto the controls.

### 2.3.2 Endogenously observed outcomes

In many labor market experiments, some outcomes of interest are observed only for a subset of individuals. For example, weekly hours of work (labor supply), hourly wages, job characteristics, career advancement, and retention on the job are observed only for those who are able to find jobs, not for those who are unemployed. Even ideal

experiments with perfect compliance may not identify the causal effects of interest on these outcomes.

### **2.3.3 Site and group effects**

Another large class of limitations in experiments has to do with generalizing beyond the experimental sample. Extrapolations to other programs, other samples, or other treatment regimes can be hazardous. We will discuss in this paper three broad classes of external validity issues.

One class has to do with variations in the treatment on offer across program locations. In many programs, the treatment is not homogeneous across locations; in other cases, the treatment may be homogeneous, but outcome distributions vary. In either case, one might be interested in identifying how treatment effects vary across locations.

The second class derives from observed differences between the population of interest and that included in the experimental sample—one might want to understand a program’s effect on a population that differs in observable ways from that represented in the experimental sample, or on a subpopulation other than the experimental compliers.

### **2.3.4 Treatment effect heterogeneity and external validity**

The third class of external validity issues arises from *unobserved* differences in individual treatment effects—when the effect of the treatment varies across individuals in ways that are not captured by observed participant characteristics, and when the parameters of interest extend beyond the ATE in the population from which the experimental sample is drawn. This can occur when, for example, the experimental complier share is not expected to match the take-up rate when the program is offered more generally, or when one expects to offer the program to a population that may differ in its treatment effect distribution from the experimental population. While conceptually similar to differences along observed characteristics, the econometrics behind addressing unobserved differences in treatment effects is sufficiently complex and self-contained that we discuss it separately.

### **2.3.5 Hidden treatments**

Interpreting estimated program effects and extrapolating to other settings can be complex even in the case of uniform treatments and uniform populations. For example, if noncompliers have access to alternatives to the program under study (e.g., to courses offered by alternative job training providers), this will lead to variation in treatment effects even without treatment effect heterogeneity or noncompliance in treatment assignment in the standard sense. The alternative treatments are often “hidden,” as administrative data on the program under study will not reveal whether participants have received alternatives elsewhere. In this case, the experimental impact identifies the treatment’s

effect relative to a poorly specified alternative that may not differ dramatically and may be a poor guide to the program's value relative to no treatment. In multisite studies, differential take-up of such hidden treatments by the control group may create the appearance of treatment effect heterogeneity across sites and hinder extrapolation to other settings.

### 2.3.6 Mechanisms and multiple treatments

In many instances, we are interested in understanding the mechanism generating a particular treatment effect. In some cases, the effects of separate mechanisms are of inherent interest. In complex experiments with multiple treatments, it is important to understand which treatments were particularly effective, and why. For example, many job training programs include JSA, and vice versa. In other cases, understanding the mechanisms is crucial in extrapolating from the particular experimental setting to other situations. For example, in the Canadian Self-Sufficiency Program (SSP), workers have to first establish eligibility to then participate a wage subsidy program, creating endogenous selection that makes it difficult to interpret how the subsidy program affects labor supply ([Card and Hyslop, 2005](#)). Without additional information or additional structure, multiple mechanisms are not separately identified, leading to potential serious limitations in understanding the program and in external validity.

## 2.4 Quasiexperimental and structural research designs

It is not always possible to use a true randomized experiment to evaluate a program or mechanism of interest, due to operational, financial, or ethical constraints. Quasiexperimental studies rely on aspects of the program or policy variation as a source of plausibly as-good-as-random variation in treatment assignment—examples include regression discontinuity designs, regression kink designs, and difference in differences (see [Angrist and Krueger, 1999](#)). These can be useful alternatives when true experiments are infeasible or simply not available. When the quasiexperimental variation is as good as randomly assigned, the various quasiexperimental designs can recover treatment effects, just as experiments can.

But even if the assumptions governing assignment are correct, quasiexperimental designs generally solve only the assignment problem and do not necessarily address the additional issues discussed above. The same is true for selection-on-observables estimators (e.g., matching estimators): the “unconfoundedness” assumption eliminates the selection problem, if it holds, but does nothing to address other design issues.

In contrast, structural approaches that explicitly specify all aspects of the choice problem and resulting outcomes can in principle resolve both assignment and other design issues simultaneously. However, this approach hinges on the model being correctly specified and hence may come at a substantial cost to internal validity.

### 3. A MORE THOROUGH OVERVIEW OF LABOR MARKET SOCIAL EXPERIMENTS

It is no accident that we discuss design issues of RCTs in the context of social experiments in the labor market, since many of the major design issues discussed in [Section 2](#) arise in the evaluation of important labor market programs. In this section we review some of the main characteristics of existing social experiments in labor economics in light of these design issues. We distinguish three broad substantive topics that have been studied extensively via social experiments: labor supply, particularly of low-income families, welfare recipients, and UI recipients; job training and skill development; and job search. In this section, we discuss each in turn. For a more detailed discussion of the experiments we mention here, we refer the reader to our summary tables and excellent overviews provided elsewhere.<sup>10</sup>

#### 3.1 Labor supply experiments

One can broadly categorize social experiments providing incentives to increase labor supply into three groups, following their program structure, target group, and time period: the IMEs in the late 1960s and early 1970s; welfare reform experiments in the late 1980s through the mid-1990s; and reemployment subsidy experiments, which span a longer time period.

##### 3.1.1 *The income maintenance experiments*

A first wave of experiments were the IMEs already discussed in [Section 2](#), which treated low-income households with various combinations of lump-sum transfers and taxes on earnings. By randomly assigning treatment and control groups to multiple treatment arms with varying combinations of tax rates and subsidies, and by separately targeting groups of different income levels, the experiments allowed the tracing out of labor supply responses in different parts of the budget constraint and under varying financial conditions. There were four such experiments, initiated between 1968 and 1971, in New Jersey, Seattle—Denver, Gary (Indiana), and in rural areas. [Table 1](#) provides detailed information about these experiments. While the sample sizes were moderate by later standards, the total cost was substantial compared to most randomized evaluation of labor supply incentives that would follow. This is in important part because the program—the payments themselves—was expensive on a per-participant basis. Complex, stratified experimental designs were used in efforts to minimize these costs, but even with these the studies were major investments.

<sup>10</sup> See among others [Greenberg and Shroder \(2004\)](#), [Heckman et al. \(1999\)](#), [Meyer \(1995\)](#). Our overview focuses almost exclusively on US experiments. For an overview of active labor market policy evaluations, drawing largely on European evidence, see [Card et al. \(2010\)](#).

Across each of the income maintenance studies and various comparison groups (e.g., husbands, wives, and single female household heads), labor supply results were fairly consistent: the combination of a lump-sum transfer and a positive tax rate reduced participants' earnings (i.e., labor supply), by more so when the transfer and tax rate were larger. This reflects a combination of income and substitution effects; [Robins \(1985\)](#) combines the various studies and uses contrasts among the different treatment arms to separately identify the income and substitution elasticities of labor supply. He concludes that these elasticities were fairly stable across studies, but fairly small: the substitution elasticity was under 0.1 for husbands, just above 0.1 for single female heads, and more variable but averaging 0.17 for wives. Income elasticities were less consistent, but centered around  $-0.1$ .

In retrospect, these experiments encountered a number of the design issues that we identified in [Section 2](#) and discuss at greater length below. For example, because of the high attrition rates, which as [Ashenfelter and Plant \(1990\)](#) note were differential across treatment groups, they also can be seen as an example of the endogenously observed outcome problem. Similarly, without additional assumptions it is impossible to estimate the effect of these programs on hours worked or wages. Interestingly, in contrast to most randomized evaluations that followed, they were primarily focused on identifying the mechanisms—income versus substitution effects—behind any labor supply responses, rather than the simple treatment effect of an existing program. This motivated the use of a large number of treatment arms, an option we discuss below as one way of addressing questions about mechanisms.

### **3.1.2 Welfare reform experiments**

A second wave of social experiments related to labor supply was initiated between the late 1980s and the mid-1990s and evaluated the effect of employment incentives for welfare recipients. While the IMEs were funded almost exclusively by the federal government, these later evaluations concerned state-level programs and were funded mostly at the state level.<sup>11</sup> In contrast to the relatively straightforward structure of the negative income tax treatments, these were usually randomized evaluations of entire, complex programs, often designed as replacements for traditional welfare, that included components designed to strengthen work incentives along with others (e.g., childcare or JSA) designed to reduce barriers to work.

We have identified welfare RCTs in at least 13 states. [Table 1](#) includes a selection of four social experiments on this topic, implemented in California, Connecticut, Florida, and Minnesota, although there were many more not listed here. A common component to most new programs (experimental treatments) was the introduction of lifetime time

<sup>11</sup> For a detailed historical account, see the chapter by Judith [Gueron \(2017\)](#) in this volume.

**Table 1** Details on selected randomized controlled trials of welfare programs and other labor supply incentives for low-income workers in the United States

	<b>Target population</b>	<b>Primary intervention</b>	<b>Secondary intervention</b>	<b>Experiment title</b>	<b>Start date</b>	<b>Cost (nominal \$)</b>
(1)	<b>Total family income not exceeding 150% of the poverty level</b>	Negative income tax		New Jersey Income Maintenance Experiment	1968	\$ 7,800,000
(2)	<b>Rural, low-income families</b>	Negative income tax		Rural Income Maintenance Experiment	1970	\$ 6,100,000
(3)	<b>Family earning less than \$11,000 in 1971 dollars</b>	Negative income tax	Vocational training	Seattle-Denver Income Maintenance Experiment	1970	\$77,500,000
(4)	<b>Black families with at least one child under the age of 18</b>	Negative income tax		Gary Income Maintenance Experiment	1971	\$20,300,000
(5)	<b>One- and two-parent families receiving AFDC</b>	Earned income disregard		California Work Pays Demonstration Program (CWPDP)	1993	\$ 4,500,000

Sample size	Treatment	Funding source	Outcomes of interest
725—Treatment 632—Control 1357—Total	Eight combinations of income guarantees and tax rates on other income.	OEO	(1) Reduction in work effort and (2) lifestyle changes
269—Treatment 318—Control 587—Total	Five negative income tax plans.	The Ford Fdn., OEO Office of Economic Opportunity	(1) Work behavior; (2) health, school, and other effects on poor children; and (3) savings and consumption behavior
1801—Treatment 1 946—Treatment 2 1012—Treatment 3 1041—Control	Two types of treatment: a negative income tax plan and a subsidy to vocational training.	HEW, HHS	(1) Effects on labor supply; (2) marital stability; and (3) other lifestyle changes
1028—Treatment 771—Control 1799—Total	Four combinations of guarantee and tax.	HEW	(1) Employment; (2) schooling; (3) infant mortality and morbidity; (4) educational achievement; and (5) housing consumption
6278—Treatment 1 3471—Treatment 2 3276—Control 1 1695—Control 2 14,720—Total	The treatment involved changing two provisions of the AFDC program. The "\$30 and one-third" provision applied to all AFDC families and allowed welfare recipients to keep the first \$30 and one-third of the remaining wages before welfare grant determinations were made. However, it expired after the recipient had been in the program for 4 months, and thereafter dollar-for-dollar reductions in grant occurred for every dollar of earnings. Under the 100-h rule, which applied only to two-parent families, the total work hours per month for the primary wage earner could not exceed 100 h without loss of eligibility. Experimental received a waiver of the time limit on the \$30 and one-third income disregard, and a waiver of the 100-h rule. However, the cash grants of experimental were reduced by 8.5%. Controls were subject to the general AFDC rules, with expiring disregards, ineligibility after 100 h, and higher benefits.	CA Department of Social Services	(1) Employment; (2) earnings; and (3) welfare receipt

*Continued*

**Table 1** Details on selected randomized controlled trials of welfare programs and other labor supply incentives for low-income workers in the United States

	<b>Target population</b>	<b>Primary intervention</b>	<b>Secondary intervention</b>	<b>Experiment title</b>	<b>Start date</b>	<b>Cost (nominal \$)</b>
(6)	<b>Families on AFDC</b>	Earned income disregard	Individual job search assistance Case management	Florida Family Transition Program (FTP)	1994	\$11,200,000
(7)	<b>AFDC recipient and recent applicant families</b>	Reemployment bonus Earned income disregard	Job search incentive Child care services	Minnesota Family Investment Program (MFIP)	1994	\$ 5,090,300
(8)	<b>AFDC recipients</b>	Earned income disregard Time limit	Job search incentives Vocational training	Connecticut Jobs First	1996	\$ 5,400,000
(9)	<b>UI claimants</b>	Reemployment bonus		Illinois Unemployment Insurance Incentive Experiment	1984	\$ 800,000
(10)	<b>UI claimants</b>	Reemployment bonus	Job search workshop	Pennsylvania Re-employment Bonus Demonstration	1988	\$ 990,000
(11)	<b>UI claimants</b>	Reemployment bonus		Washington State Reemployment Bonus Experiment	1988	\$ 450,000

DOL, US Department of Labor; ETA, Employment and Training Administration; Fdn., Foundation; OEO, Office of Economic Opportunity; HEW, US Department of Health, Education, and Welfare; HHS, US Department of Health and Human Services.

(1) Kershaw and Fair, 1976; Watts and Rees, 1977a,1977b; (2) US Department of Health, Education, and Welfare 1976; Palmer and Pechman, 1978; (3) SRI International, 1983; (4) Kehrer, 1977; (5) Becerra et al., 1998; (6) Bloom et al., 2000; (7) Knox et al., 2000; (8) Bloom et al., 2002; (9) Woodbury and Spiegelman, 1987; (10) Corson et al., 1991; (11) Spiegelman et al., 1992.

Sample size	Treatment	Funding source	Outcomes of interest
1400—Treatment 1400—Control 2800—Total	Limited welfare benefits unless “job-ready,” enhanced earnings disregard, and intensive case management.	FL Department of Children and Families; US Department of Health and Human Services	(1) Earnings; (2) welfare benefit receipts; and (3) outcomes or children
5275—Treatment 1 1933—Treatment 2 5634—Treatment 3 1797—Control 14,639—Total	MFIP provided a 20% grant increase when recipients became employed, increased the level of income that would be disregarded in grant calculation, and paid the child care subsidy directly to caregiver. Two-parent families were not subject to work history requirements or to the 100-h rule. Both single-parent and two-parent families assigned to MFIP were subject to mandatory participation in employment services. Rules and procedures were simplified by combining Food Stamps, AFDC, and Minnesota’s Family General Assistance (FGA) to form a single cash benefit program. Subjects assigned to the MFIP incentives-only group received identical benefits as MFIP, but were not required to participate in training services. Two other groups.	MN Department of Human Services; Ford Fdn.; HHS; US Department of Agriculture; Charles Stewart Mott Fdn.; Annie E Casey Fdn.; McKnight Fdn.; Northwest Area Fdn.	(1) Employment; (2) earnings; (3) welfare receipt; (4) total family income; and (5) other measures of child and family well-being
2138—Treatment 1821—Control 3959—Total	Earnings disregarded below the federal poverty level and required to participate in Job Search Skills Training.	CT Department of Social Services	(1) Employment; (2) earnings; (3) benefit receipt; and (4) other measures of child well-being
4186—Treatment (claimants) 3963—Treatment (employers) 3963—Control 12,112—Total	Unemployed were offered a \$500 bonus if found a job within 11 weeks and held it for 4 months.	IL Department of Employment Security; WE Upjohn Institute for Employment Research	(1) Reductions in unemployment spells and (2) net program savings
14,086—Treatment 3392—Control 17,478—Total	Five combinations of bonus amount and qualification period.	DOL	(1) UI receipt; (2) employment; and (3) earnings
12,451—Treatment 3083—Control 15,534—Total	Six variations of reemployment bonus amount and qualification periods.	Alfred P Sloan Fdn. US DOL, ETA	(1) Weeks of insured unemployment and (2) UI receipt

limits of welfare receipt and increases in earnings disregards, both eventual components of the 1996 federal welfare reform—prior to this reform, implementation of such changes required a waiver from the US Department of Health and Human Services, and this was often conditioned on an experimental evaluation. The exact nature of both the new programs and the traditional welfare benefit varied by state. Other program features varied widely as well, including JSA, access to childcare, changes in case management, and provision of job training.

Two examples to which we will refer to later are Connecticut's Jobs First and Florida's Family Transition Program. In both cases, control group members faced a welfare benefit schedule that had no time limits and high implicit taxes on working.<sup>12</sup> Jobs First and the Family Transition Program each introduced time limits for welfare receipt and benefit schedules with lower implicit tax rates. Under Jobs First, eligible welfare recipients saw no reduction in their benefits while working until earnings hit the federal poverty line. Under the Family Transition Program, a working welfare recipient could keep \$200 a month, plus 50% of all earnings above \$200. Both programs also modified other welfare program features, including enhanced enforcement of work requirements, changing the duration of access to Medicaid benefits, setting asset limits for welfare receipt, and providing childcare assistance, among others.

The randomized evaluation of the two programs captured the combined effects of all of these changes on employment and earnings. Each program led to higher earnings and higher total incomes, inclusive of welfare payments, in the treatment group, though in each case this effect diminished over time. Total governmental costs were higher for the Connecticut treatment group than for controls, but the reverse was true in Florida. An important caveat is that these results largely reflect the period before time limits bound.

In many of the welfare-to-work experiments, key outcomes of interest included hours of work among those who are employed and wages or earnings. Neither of these is observed for those who are not employed. Thus, although many studies report experimental effects on endogenously observed outcomes, these are understood to suffer from serious selection problems. Another issue to take into account in interpreting these experiments is the possibility of spillover effects. These were typically not small pilot studies but involved broad changes to welfare rules, sometimes applied to all program participants except for a hold-out control group.

<sup>12</sup> In Connecticut, welfare recipients were eligible for a fixed earnings disregard of \$120 for the 12 months following the first month of employment while on assistance and \$90 afterward. Recipients were also eligible for a proportional disregard of earnings above \$120 (\$90): 51% for the 4 months following the first month of employment and 27% afterward. In Florida, after the first 4 months of work, the marginal tax rate on earnings for AFDC recipients was 100% if they earned over \$90 per month.

Another major question regarding welfare-to-work programs concerns heterogeneity in treatment effects. One might imagine that there is a subpopulation of recipients who are responsive to work incentives and another group of hard cases who are much less responsive. The ATEs that can be estimated from these experiments might substantially overstate the employability of the latter participants.

### **3.1.3 Reemployment subsidy experiments**

A third broad group of labor supply-related experiments evaluated direct reemployment subsidies. One set of such programs had incentives structured like a negative income tax and were targeted to welfare recipients or low-income individuals, sometimes as part of the same AFDC reforms discussed above. These took place mostly in the mid- to late 1990s and included the Canadian SSP, Minnesota's Family Investment Program (FIP), and Wisconsin's New Hope Project. These RCTs can be seen as evaluations of welfare-like programs, but included subsidies that were conditional on sustaining a certain amount of employment. Not surprisingly, these programs generally led to increased earnings among treatment group participants (although FIP was an exception); different studies varied in whether the additional income of participants was larger or smaller than the extra welfare costs borne by the government.

Another set of such programs were schemes that paid lump-sum subsidies conditional on employment—effectively, bonuses for finding work. These include the well-known reemployment bonus experiments targeted at unemployed workers receiving UI in Illinois, Pennsylvania, and Washington State in the mid-1980s. These studies found that eligibility for a relatively large reemployment bonus led to shorter UI spells, with no detectable impact on the quality of the job obtained, but that the effects were relatively small and thus the programs were not cost-effective.

More recently, a bonus for welfare recipients who found a job and who remained reemployed for a certain time was evaluated in the context of Texas' Employment Retention and Advancement (ERA) project in the early 2000s ([Dorsett et al., 2013](#)). The Texas evaluation was part of a large-scale randomized evaluation of 12 different service combinations in different US cities from 2000 to 2004 under the ERA project umbrella ([Hamilton and Scrivener, 2012](#)). The main focus of ERA was to expand workforce services to recently reemployed welfare recipients or low-wage workers to maintain successful labor force attachment (though three sites, including Texas, combined pre- and postemployment assistance). The evaluation tested a broad range of services, with at best mixed results regarding the effect of postemployment services tested.

An important feature of several of these employment subsidy programs was that potential recipients had to become eligible for the subsidy, usually by working a minimum amount of hours. Hence, while the main goal of the programs was to help workers build attachment to the labor force, effects of the subsidy (as distinct from the subsidy offer) on the duration of employment could be estimated only for those who found

jobs in the first place, a subsample that was differentially selected in the treatment and control groups. [Card and Hyslop \(2005\)](#) refer to this as an “eligibility effect”; in our earlier taxonomy of design challenges, this can be seen as a case where the mechanisms underlying the treatment effect are of primary interest. Under any name, it complicates the interpretation of the outcomes of a simple RCT.

Overall, randomized studies of a range of labor supply incentive programs have found labor supply responses to changes in implicit or explicit financial incentives as predicted by theory. However, a broad theme emerges that employment effects have mostly been short-lived and effects on total participant income have been inconsistent. A challenge in interpreting these studies has been that typically a number of treatments were varied simultaneously, including implicit tax rates and lump-sum transfers, training programs, JSA, enforcement, and/or time limits. Hence, extrapolating from these findings to new programs providing different combinations of treatments is difficult without understanding the underlying behavioral responses, which typically requires additional assumptions.

### **3.2 Training experiments**

From 1964 to today, we count over 50 RCTs that evaluate job training programs of various forms. These include large-scale evaluations conducted at the national level, state-level evaluations, and evaluations of programs at the local level. The programs evaluated varied substantially in the type of training, which ranged from vocational and general classroom-based training of different durations to on-the-job training by actual employers. Most training programs were complemented by some kind of JSA, but in the studies we review here this was not the emphasis. [Table 2](#) provides an overview of a selected group of these RCTs.

Training programs are less easily classified than labor supply programs. While the first job training social experiment of which we are aware focused on laid off workers (the General Education in Manpower Training experiment, begun in 1964), the vast majority of training programs are targeted to welfare recipients, to low-income individuals generally, or to low-income youth. Moreover, while one can broadly distinguish phases of experimental evaluation parallel to the patterns in the evaluation of welfare programs outlined above, randomized evaluations of training programs occurred more evenly from the 1980s to today. It is also harder to discern common patterns in the types of training provided or programs evaluated.

The first large-scale evaluation of a mix of on-the-job experience and supervision for hard-to-employ individuals was the National Supported Work Demonstration (NSWD), which ran from 1975 to 1980. The NSWD was a large and expensive social experiment implemented by the United States at the national level, but did not evaluate an established training program. Rather, the NSWD relied on local nonprofits to

**Table 2** Details on selected randomized controlled trials of programs offering job training and work experience for low-income individuals in the United States

Target population	Primary intervention	Secondary intervention	Experiment title	Start date	Cost (nominal \$)	Sample size	Treatment	Funding source	Outcomes of interest
(1) AFDC recipients, ex-offenders, substance abusers, and high school dropouts	Work experience		National Supported Work Demonstration (NSWD)	1975	\$ 82,400,000	3214—Treatment 3402—Control 6616—Total	Employment in a structured work experience program involving peer group support, a graduated increase in work standards, and close sympathetic supervision, for 12–18 months.	DOL, ETA; DOJ; Law Enforcement Assistance Administration; HHS; National Institute on Drug Abuse; HUD; US Department of Commerce; Ford Fdn.	(1) Increases in posttreatment earnings; (2) reductions in criminal activity; (3) reductions in transfer payments; and (4) reductions in drug abuse
(2) AFDC recipients	Work experience		AFDC Homemaker—Home Health Aide Demonstrations	1983	\$ 8,000,000	4750—Treatment 4750—Control 9500—Total	Experimental AFDC subjects (trainees) received a 4- to 8-week training course to become a homemaker-home health aide, followed by a year of subsidized employment. Control subjects did not receive this training, nor did they receive subsidized employment.	Health Care Financing Administration	(1) Employment; (2) earnings; and (3) AFDC and food stamp payments and receipt
(3) Eligible Job Training Partnership Act Title II adults and out-of-school youth	Vocational training General education Work experience On-the-job-training	Individual job search assistance	National Job Training Partnership Act (JTPA) Study	1987	\$ 23,000,000	20,602	Classroom training, on-the-job training, job search assistance, basic education, and work experience.	DOL	(1) Earnings; (2) employment; (3) welfare receipt; and (4) attainment of educational credentials and occupational competencies
(4) AFDC recipients	Vocational training General education Work experience	Individual job search assistance	Greater Avenues for Independence (GAIN)	1988		24,528—Treatment 8223—Control 32751—Total	Basic education, job search activities, assessments, skills training, and work experience.	California Department of Social Services (CDSS)	(1) Participation in employment-related activities; (2) earnings; (3) welfare receipt; and (4) employment
(5) All recipients of ADC (Ohio's AFDC program)	Work experience General education	Individual job search assistance	JOBS	1989	\$ 3,000,000	24,120—Treatment 4371—Control	Mandatory employment and training services, which included basic and post secondary education, community work experience, and job search assistance.	OH Department of Human Services	(1) Employment; (2) earnings; and (3) welfare receipt

*Continued*

**Table 2** Details on selected randomized controlled trials of programs offering job training and work experience for low-income individuals in the United States—cont'd

Target population	Primary intervention	Secondary intervention	Experiment title	Start date	Cost (nominal \$)	Sample size	Treatment	Funding source	Outcomes of interest
(6) <b>Low-income, disadvantaged workers and job seekers</b>	Vocational training	Individual job search assistance	Sectoral Employment Impact Study	2003		1286—Total	Industry-specific training programs that prepared unemployed and underskilled workers for skilled positions and connected them with employers seeking to fill such vacancies. Sectoral programs employ various approaches depending on the organization leading the effort and local employers' needs.	Charles Stewart Mott Fdn.	(1) Earnings; (2) employment; and (3) quality of jobs
(7) <b>Low-wage workers</b>	Vocational training On-the-job training	Case management	Work Advancement and Support Center (WASC) Demonstration	2005		1176—Dayton 971—San Diego 705—Bridgeport 2852—Total	The program offered participating workers intensive employment retention and advancement services, including career coaching and access to skills training. It also offered them easier access to work supports, in an effort to increase their incomes in the short run and help stabilize their employment. Finally, both services were offered in one location—in existing One-Stop Career Centers created by the Workforce Investment Act (WIA) of 1998—and by colocated teams of workforce and welfare staff.	State of Ohio; County of San Diego Health and Human Services Agency; DOL ETA; U.S. Department of Agriculture, Food and Nutrition Service; HHS; Administration for Children and Families; Ford Fdn.; Rockefeller Fdn.; Annie E. Casey Fdn.; David and Lucile Packard Fdn.; The William and Flora Hewlett Fdn.; Joyce Fdn.; James Irvine Fdn.; Charles Stewart Mott Fdn.; Robert Wood Johnson Fdn.	(1) Employment and (2) earnings (along with many other outcome measures)

(8)	<b>School dropouts aged 17–21 years</b>	General education Vocational training	Individual job search assistance	JOBSTART	1985	\$ 6,200,000	1163—Treatment 1149—Control 2312—Total	Education and vocational training, support services and job placement assistance.	DOL; Rockefeller Fdn.; Ford Fdn.; Charles Stewart Mott Fdn.; William and Flora Hewlett Fdn.; more foundations.	(1) Educational attainment; (2) employment; (3) earnings; and (4) welfare receipt
(9)	<b>16–24-year olds</b>	General education Vocational training	Health care services Housing services	National Job Corps Study	1994	\$ 21,587,202	9409—Treatment 5977—Control 15,386—Total	Treatment group allowed to enroll in Job Corps group. Job Corps centers provide vocational training, academic instruction, health-care, social skills training, and counseling.	DOL, ETA	(1) Employment; (2) earnings; (3) education and job training; (4) welfare receipt; (5) criminal behavior; (6) drug use; (7) health factors; and (8) household status

*DOJ*, US Department of Justice; *HHS*, US Department of Health and Human Services; *HUD*, US Department of Housing and Urban Development; *DOL*, US Department of Labor; *ETA*, Employment and Training Administration; *Fdn.*, Foundation.

(1) [MDRC Board of Directors, 1980](#); (2) [Bell et al., 1987](#); (3) [Bell et al., 1994](#), [Bloom et al., 1997](#); (4) [Freedman et al., 1994](#); (5) [Fein et al., 1994](#); (6) [Maguire et al., 2010](#); (7) [Miller et al., 2012](#); (8) [Cave et al., 1993](#); (9) [Burghardt et al., 2001](#).

organize a program in which treatment participants were placed in teams of up to 10 participants working under a foreman, who also served as a counselor and later provided JSA, on small-scale projects, typically in construction, light manufacturing, or social service provision. Participants received as much as 1 year of work experience, under conditions of increasing demands, close supervision, and work in association with a crew of peers. The study targeted four groups of workers: women that had been on AFDC for at least 30 months; ex-addicts; ex-offenders; and young high-school dropouts. It took place at 10 sites, and at each sites, enrollees were selected randomly from a group of volunteers.<sup>13</sup> Participation had large positive effects on AFDC recipients and smaller positive effects on ex-addicts, but benefits for other groups were smaller and generally statistically insignificant.

The data used to evaluate NSWWD came from a series of follow-up surveys.<sup>14</sup> Attrition was an issue here: after 27 months, only 72% (68%) of the treatment (control) groups of the NSWWD completed interviews. As in the Negative Income Tax (NIT) studies, this can be seen as a variant of the endogenously observed outcomes problem.

The NSWWD study was followed by a range of evaluations of state-level programs in the early- to mid-1980s. These were targeted almost exclusively at welfare recipients and largely financed by the federal government. These evaluations continued, with greater involvement of state governments, through the late 1980s and mid-1990s. While many of these RCTs were relatively small, some were substantial. Examples include the California GAIN and Ohio JOBS program evaluations, beginning in 1988 and 1989, respectively. Detailed characteristics of some of these evaluations are shown in [Table 2](#). The California program, which was mandatory for welfare recipients, included JSA, basic education, and skills training. It had large positive effects on earnings and negative effects on welfare receipt, particularly for single parents. Effects were largest in Riverside County, where administrators emphasized job placement as the central goal. However, a reanalysis of the long-term effects of GAIN by [Hotz et al. \(2006\)](#) found that the effects in Riverside County were short-lived relative to those in Los Angeles County, which focused more on human capital development and where effects were initially smaller but rose over time.<sup>15</sup> The Ohio program was similar in design but encountered more problems in implementation and yielded smaller effects.

<sup>13</sup> The Manpower Demonstration Research Corporation (MDRC) was founded in 1974 to manage the NSWWD study. For a detailed summary of the program and findings, see [Manpower Demonstration Research Corporation Board of Directors \(1980\)](#).

<sup>14</sup> The NSWWD has been examined by an extensive literature, including [LaLonde \(1986\)](#), [Dehejia and Wahba \(2002\)](#), and [Smith and Todd \(2005\)](#).

<sup>15</sup> [Hotz et al. \(2006\)](#) also point out that the treatment group was selected differently between the four GAIN sites, possibly contributing to the estimated “site” effects. For example, the Riverside County RCT sample included a smaller fraction of the more disadvantaged welfare recipients.

An exception to the trend toward evaluation of state-level or local training programs was the large-scale, national evaluation of the main federal training program aimed at low-income adults and disadvantaged youth—the National Job Training Partnership Act (JTPA) Study. The JTPA was a federal program enacted in 1982 and was administered at the state and local level. JTPA training programs provided employment training for specific occupations and services, such as JSA and remedial education, to roughly 1 million economically disadvantaged individuals per year. While the program and some services were administered directly by JTPA staff, training was provided through local service providers, such as vocational—technical high schools, community colleges, proprietary schools, and community-based organizations. Training lasted 3–4 months, on average, but duration varied widely across individuals and program sites.

Congress, in part responding to limitations of nonexperimental evaluations of the predecessor program to JTPA, the Comprehensive Employment and Training Act, mandated a randomized evaluation of JTPA in 1986. Control subjects were excluded from obtaining JTPA services for 18 months. To assess short- and medium-term program impacts on employment and earnings, the evaluation both collected survey data and drew from administrative state-level records.<sup>16</sup> The evaluation took place at 16 JTPA program sites (so-called Service Delivery Areas, SDAs). Participation by SDAs in the evaluation was voluntary, and some SDAs objected to randomly excluding eligible applicants. The participating SDAs did not differ from others in observable characteristics (e.g., Bloom et al., 1997), but may have differed in unobserved ways that would be relevant to an extrapolation to the overall effect of the national program.

An explicit goal of the JTPA evaluation was to obtain differential impacts for a wide range of target groups, including adult women, adult men, female youths, and male youth with and without an arrest record. Adult women saw the largest earnings gains, followed by adult men; effects on youth were smaller and generally not significant (although there were significant effects on attainment of high-school diplomas for both adult women and female youth). In addition to demographic subgroup analyses, heterogeneity in program impacts was estimated along several other dimensions, including JTPA services recommended by program intake staff, ethnicity, and prior labor market experience. While the subgroup effects of interest were largely prespecified, this does not fully eliminate multiple-comparisons problems, particularly when the number of prespecified comparisons is so large, and thus there is an enhanced risk of a false positive.

Job training evaluations slowed after welfare reform in the mid-1990s and then began to pick up again in the early 2000s. Some evaluations in this period focused on

<sup>16</sup> See Bell et al. (1994) and Bloom et al. (1997) for descriptions of the JTPA evaluation. There is a substantial literature on the evaluation of the JTPA program. See Heckman et al. (1999) for a summary.

sector-specific employment, such as the Sectoral Employment Impact Study (e.g., [Maguire et al. \(2010\)](#)) and evaluations of similar smaller, local programs.<sup>17</sup> There was also a randomized evaluation of combined training and job placement services under the Workforce Investment Act (WIA) from 2005 to 2015 (the Work Advancement and Support Center Demonstration), and more recently a study of the return from community college attendance under the Trade Adjustment Assistance Community College and Career Training (TAACCCT) Grants Program.

A distinct broad strand of randomized evaluations of training programs focuses on low-income youths. Again, these programs offer a broad range of different types of training augmented by varying combinations of support services. Social experiments in this area have included a range of federally and nationally funded evaluations ranging from the early 1980s to the mid-1990s that culminated in the National Jobs Corps Study, described below. As in other job training studies, the pace of experimentation slowed in the mid-1990s, but several new studies were undertaken in the mid-2000s. Some randomized evaluations, such as New York City's Summer Youth Employment Program (strictly, a natural experiment, as randomization is part of the rationing process and not a decision made to facilitate an evaluation), are ongoing. Again, the broad trend was from a federal monopoly on funding toward a greater involvement of local and private funding sources.

The largest and perhaps best known study of a training program for disadvantaged youths is the National Jobs Corps Study. The Job Corps was created in 1964 as part of the War on Poverty and currently operates under the provisions of the Workforce Innovation and Opportunity Act of 2013, which consolidated programs authorized under the WIA of 1998. Job Corps services are geared toward economically disadvantaged youths aged 16 to 24. Core services are delivered by a Job Corps center, usually residential, and include vocational training, academic education, residential living, health care, and a wide range of other services, including counseling, social skills training, health education, and recreation.<sup>18</sup> About a quarter of the over 100 centers are operated directly by the US government, with the remainder operated by private contractors. The average duration of the program is 8 months, although by its philosophy the duration responds to the participant's needs and actual duration varies widely. For 6 months after the youths leave the program, placement agencies help participants find jobs or pursue additional training.

<sup>17</sup> These include, among others, the Georgia Works programs, Project Quest in San Antonio, the Wisconsin Regional Training Partnership in Milwaukee, Per Scholas in New York City, and the Jewish Vocational Service in Boston.

<sup>18</sup> The majority of training is vocational, and curricula were developed with input from business and labor organizations and emphasize the achievement of specific competencies necessary to work in a trade. Academic education aims to alleviate deficits in reading, math, and writing skills and to provide a GED certificate. Although most Job Corps services are residential, there have been nonresidential participants (mostly women with children). There have been efforts to evaluate nonresidential Job Corps services (e.g., [Greenberg and Shroder, 2004](#); [Schochet et al., 2008](#)).

The Job Corps evaluation was based on an experimental design in which, with a few exceptions, all youths nationwide who applied to Job Corps in the 48 contiguous states between November 1994 and December 1996 and were found to be eligible were randomly assigned to either a program group or a control group. Program group members were allowed to enroll in Job Corps; control group members were excluded for 3 years after random assignment. The comparisons of program and control group outcomes represent the effects of Job Corps relative to other available programs that the study population would enroll in if Job Corps were not an option.<sup>19</sup> The control and treatment groups were tracked with a series of interviews immediately after randomization and continuing 12, 30, and 48 months after randomization.

The evaluation of Job Corps followed the outcomes of over 15,000 experimental subjects for up to 8 years using survey and administrative data. The effect of training on earnings became gradually positive as individuals graduated from the program and then remained statistically significantly different from the control group for up to 4 years afterward. At the same time, government transfers and crime rates fell (e.g., [Schochet et al., 2008](#)). There was substantial heterogeneity in outcomes—the effects were strongest for those 20- to 24-year old at the time of training and weakest for Hispanics.

A concern with these findings was that the overall level of earnings and the size of the treatment effects were quite different in the administrative data than in the survey data. While survey data are more likely to be affected by endogenous attrition, administrative data are not a panacea: they exclude under-the-table employment, which may be common in the Job Corps population.<sup>20</sup> They also cannot address the problem that wages are observed only for those who are employed, itself an intermediate outcome of the program (e.g., [Lee, 2009](#)).

An important question regarding Job Corps is the relative performance of the different Job Corps centers, which operate in different labor markets and are (sometimes) run by contractors rather than directly by the government. [Schochet and Burghardt \(2008\)](#) use the Job Corps evaluation data to estimate separate treatment effects by site, finding that these are not strongly correlated with the nonexperimental measures that have been used to assess site performance.

A final issue in the Job Corps evaluation, not to our knowledge addressed in the literature, is that the program may be large relative to the relevant labor markets, creating the possibility of important spillovers from treated to control study participants.

<sup>19</sup> Of course, if Job Corps did not exist, the ecosystem of other available programs would presumably change. This is formally a SUTVA violation and implies that control group mean outcomes may not equal what would be seen in the absence of the program.

<sup>20</sup> [Kornfeld and Bloom \(1999\)](#) show that this is the case for participants in the Job Training Partnership Act (JTPA) evaluation.

A final, smaller category of large-scale social experiments of training programs focused specifically on unemployed (displaced) workers. As we will discuss below, some of these RCTs evaluated programs providing a broad array of reemployment services that also included some degree of training. This raises a similar issue to what we highlighted above with welfare experiments—experimental evaluations generally identify the “black box” effect of the overall programs, but not the components or mechanisms responsible for those effects.

The Individual Training Account (ITA) experiment running from 2001 to 2005 directly evaluated different modes of training provision prescribed by the 1998 WIA. WIA allowed local agencies to impose different degrees of counseling and supervision of workers’ training choices, and the ITA experiment evaluated the effect of these choices on actual training received and labor market outcomes. Effectively, the ITA experiment compared three service models. Guided Choice and Maximum Choice had standardized subsidies for training, but the former required counseling by a case worker while the latter had no counseling requirement. A third model, Structured Choice, was effectively like Guided Choice but offered individualized, and typically more generous, training awards.<sup>21</sup>

The findings indicated that either more generous awards (Structured Choice) or less counseling (Maximum Choice) led to a higher incidence of training ([Perez-Johnson et al., 2011](#)). Earnings increased for workers in Structured Choice relative to Guided Choice 5 years after the treatment. (Earnings effects were higher but not statistically different for Maximum Choice relative to Guided Choice or to a control group.) While Structured Choice was estimated to be cost efficient to society, it was more expensive for the workforce system, and most agencies adopted Guided Choice as the leading model. More recently, an ongoing experiment (the WIA Adult and Dislocated Worker Programs Gold Standard Evaluation, discussed below) evaluates directly the intensive and training services provided under WIA.

An issue that is common to all of the job training experiments is the possibility that individuals assigned to the control group may have received training through other channels that would not necessarily have been tracked in the experimental data. These hidden treatments are likely to attenuate the estimated training effects—insofar as control participants are receiving substitute treatments, the evaluations identify only the *differential* effect of the public training program, rather than the overall effect of training relative to none. While this could partly explain low estimated treatment effects, this has not been examined carefully in the literature (although, as we discuss below, it has received

<sup>21</sup> Originally, under Structured Choice, caseworkers were supposed to play a more active role in training choice. However, most caseworkers did not feel they had enough knowledge of local labor markets or the worker’s skills to take on such an active role.

substantial attention in some other domains, most notably the evaluation of early childhood education).

Although a broad range of findings from different treatments make it hard to generalize, two themes have emerged from training program social experiments. First, while training for less advantaged adults and the unemployed can have beneficial effects, most training programs for disadvantaged youths fail to achieve strong results. An important exception is Job Corps, which has shown short- and medium-term positive effects for at least some of its participants. Second, the effects of training tend to accrue gradually over time, making them hard to detect in research designs that combine multiple treatments or that do not have sufficient data or samples to precisely estimate medium- to long-term effects.

### 3.3 Job search assistance

From the inception of welfare programs in the United States, it was suspected that neither better work incentives nor better human capital would be sufficient to place hard-to-employ welfare recipients or disadvantaged youth into lasting employment and that part of the challenge derived from disconnection from the world of work. At the same time, it was not clear which range of support services aiding job placement would be effective. Hence, a large number of RCTs have evaluated a range of JSA programs for low-income workers and youth ([Table 3](#)). Other studies have focused on UI recipients and other unemployed workers, who have traditionally been eligible for search assistance from the US government. Hence, while training evaluations have mostly concerned programs aimed at low-income workers, JSA experiments have evaluated programs geared toward a wider range of unemployed workers from the mid-1970s to today. As in training evaluations, however, an important challenge in studies of JSA is measuring the counterfactual: what sort of assistance, if any, was received by those excluded from the program under study?

An early wave of JSA program experiments geared toward welfare recipients occurred from the early 1970s to the mid-1980s, alongside similar studies of labor supply and training programs aimed at the same population. These were mostly evaluations of local programs funded by the federal government. There is a long history of programs providing placement and training services for welfare recipients in the United States, going back at least to the Work Incentive Program (WIN) initiated in 1967. WIN was criticized on a range of fronts (e.g., [Gold, 1971](#)). The first wave of federally funded evaluations tested services provided by the WIN program and alternative programs for WIN-eligible welfare recipients (e.g., [Grossman and Roberts, 1989](#)). These culminated in the National Evaluation of Welfare-to-Work Strategies (NEWWS) in 1990, which was a large-scale evaluation of 11 programs combining JSA, training, and enforcement of job search requirements in seven different sites in the United States.

**Table 3** Details on selected randomized controlled trials of job search assistance programs for low-income individuals and unemployed workers in the United States

Target population	Primary intervention	Secondary interventions	Experiment title	Start date	Cost (nominal \$)	Sample size	Treatment	Funding source	Outcomes of interest
(1) Single-parent heads of household who were required to participate in the program (recipients of AFDC)	Job Club	General education Vocational training	Project Independence—Florida	1990	\$ 3,600,000	13,513—Treatment 4274—Control 17,787—Total	The experimental group was eligible to receive Project Independence services and was subject to a participation mandate. Services included independent job search, job club, assessment, basic education, and training. The control group was not eligible for these services and was not subject to a participation mandate.	Florida Department of Health and Rehabilitative Services Ford Fdn. US Department of Health and Human Services	(1) Employment; (2) earnings; and (3) AFDC receipt
(2) Single-parent welfare recipients	Job Club Case Management	General education Vocational training	National Evaluation of Welfare-to-Work Strategies (NEWWS)	1991	\$ 31,700,000	44,569—Total	Eleven programs, broadly defined as either employment-focused or education-focused, were tested in seven sites across the United States.		(1) Employment; (2) earnings; (3) welfare receipt; (4) cost-effectiveness; and (5) child well-being
(3) Families on welfare	Individual job search assistance	Earned income disregard Work experience	Indiana Welfare Reform Evaluation	1995	\$ 23,200,000	63,223—Treatment 1 3863—Treatment 2 3217—Control 1 1091—Control 2 71,394—Total	Experimentals were subject to new welfare reform policies: assisted job search, broader mandatory work participation, earned income disregard, time limits for case assistance, a revised system of child care provision, family benefit cap, and parental responsibility (such as immunizing children). Controls continued under the traditional AFDC policies.	Indiana Family and Social Services Administration US Department of Health and Human Services	(1) Employment; (2) earnings; (3) welfare receipt; (4) income; (5) health insurance; and (6) parental responsibility

(4)	<b>Single-parent (AFDC-FG) and two-parent (AFDC-U) welfare families in Los Angeles County</b>	Job Club Individual job search assistance  Job search workshop	LA Jobs-First GAIN Evaluation	1995	\$ 29,900,000	11,521—Treatment 1 4039—Treatment 2 4162—Control 1 1009—Control 2 20,731—Total	<p>Members of the treatment group were enrolled in Jobs-First GAIN. These subjects were required to participate in at least one of the job search activities, including job clubs and other informational services and job search training sessions. Experimental units were also exposed to Jobs-First GAIN's intensive work-first message. Sanctions were imposed, usually in the form of partial reductions in welfare benefits, for failure to participate. Controls were not exposed to any of Jobs-First GAIN's services, the intensive work-first message, or sanctions. Controls could still receive assistance from other agencies and were subject to existing welfare rules.</p>	Los Angeles Department of Public Social Services US Department of Health and Human Services Ford Fdn.	(1) Employment; (2) earnings; (3) welfare benefits; (4) outcomes for children; and (5) incremental effects compared with previous LA GAIN program
(5)	<b>UI claimants</b>	Individual job search assistance	Vocational training	Nevada Claimant Placement Program (NCPP)	1977	3500	More staff attention and more referrals, weekly interviews and eligibility checks, all services from same ES/UI team which coordinated their efforts.		(1) Weeks of benefits; (2) earnings; (3) enforcement of work search rules; (4) job searches; and (5) referrals and placements

*Continued*

**Table 3** Details on selected randomized controlled trials of job search assistance programs for low-income individuals and unemployed workers in the United States—cont'd

Target population	Primary intervention	Secondary interventions	Experiment title	Start date	Cost (nominal \$)	Sample size	Treatment	Funding source	Outcomes of interest
(6) UI claimants	Job search incentives Individual job search assistance		Claimant Placement and Work Test Demonstration	1983	\$ 225,000	1485—Treatment 1 1493—Treatment 2 1666—Treatment 3 1277—Treatment 4	Job search and placement services.	US Department of Health and Human Services; Ford Fdn.	(1) Employment and (2) UI payments reductions
(7) UI claimants indefinitely separated from most recent job	Individual job search assistance		Wisconsin Eligibility Review Pilot Project (ERP)	1983		5000	6-h job search workshop conducted by ES staff; also tried 3-h job search workshop.		(1) Weeks of benefits; (2) earnings; (3) enforcement of work search rules; (4) job searches; and (5) referrals and placements
(8) Unemployed	Case management Individual job search assistance Job search workshop		Reemploy Minnesota (REM)	1988	\$ 835,000	4212—Treatment Unknown—Control (roughly 10 times treatment)	More personalized and intensive unemployment insurance (UI) services, including case management, intensive job search assistance and job matching, claimant targeting for special assistance, and a job-seeking skills seminar. The control group received regular UI services.	Unemployment Insurance Contingent Account of the Minnesota Department of Jobs and Training	(1) Duration of UI benefits and (2) amount of UI benefits
(9) UI claimants	Individual job search assistance	Vocational training	Kentucky Worker Profiling and Reemployment Services (WPRS) Experiment	1994	\$ 15,000	1236—Treatment 745—Control 1981—Total	Structured job search activities, employment counseling, and retraining.	Kentucky Department of Employment Services	(1) Earnings; (2) length of benefit receipt; and (3) amount of UI benefits received

(10)	<b>UI claimants</b>	Alternative work search policies		Maryland Unemployment Insurance Work Search Demonstration	1994	\$ 250,000	3510—Treatment 1 3455—Treatment 2 3680—Treatment 3 3400—Treatment 4 4812—Control 1 4901—Control 2 23,758—Total	Four different rules changes to Maryland UI eligibility rules.	US DOL ETA	(1) UI payments in terms of weeks and dollars; (2) continuing eligibility; (3) employment; and (4) earnings
(11)	<b>UI claimants</b>	Individual job search assistance Case management	Vocational training	Reemployment and Eligibility Assessment (REA)	2013			(1) Current REA Program: Assistance—defined as the provision of labor market information, developing an individual reemployment plan, a referral to reemployment services and direct provision of reemployment services + enforcement (see below) (2) Enforcement Only: The requirement that claimants appear for the REA meeting and that REA program staff verify claimants' eligibility and their participation in work search activities with referral to adjudication and possible suspension of UI benefits for those who do not participate.	US DOL ETA	(1) UI benefit receipt; (2) employment; and (3) earnings

DOL, US Department of Labor; ETA, Employment and Training Administration; Fdn., Foundation.

(1) [Kemple et al., 1995](#); (2) [Hamilton et al., 2001](#); (3) [Beecroft et al., 2003](#); (4) [Freedman et al., 2000](#); (5) [Steinman, 1978](#); (6) [Johnson et al., 1984](#); [Corson et al., 1984](#); (7) [Herrem and Schmidt, 1983](#); [Jaggers, 1984](#) (8) [Minnesota Department of Jobs and Training, 1990](#); (9) [Black et al., 2003](#); (10) [Klepinger et al., 1997](#); (11) [Klerman et al., 2013](#).

The results from randomized evaluation of different WIN services were mixed (e.g., [Greenberg and Shroder, 2004](#)). The evaluation of so-called “job clubs” in 1976–79 showed substantial increases in employment and reduction in welfare receipt. As a result, job clubs became an integral part of services received by welfare recipients. However, the evaluation was based on a relatively small sample, follow-up was limited to 1 year, and the results indicated substantial, hard-to-explain heterogeneity in the findings across subgroups and treatment sites. In contrast, the evaluations discussed in [Grossman and Roberts \(1989\)](#) show less consistent effects of JSA under the WIN program.

The much larger evaluation of NEWWS found short-term increases in employment and reductions in welfare receipt. These effects dissipated during the 5-year follow-up period. As in other evaluations occurring in the early to mid-1990s, such as GAIN discussed above, this may be due in part to the high-pressure labor market of the 1990s. The presence of such cyclical effects is a potentially important confounder limiting the interpretation of the effects of labor market program studies.

A second wave of experiments occurred in the run-up to welfare reform in the mid-1990s and again saw substantial state-level involvement. As with labor supply and training studies in this period, these studies tended to study contemplated changes to existing programs and to involve large samples. These included Project Independence in Florida in 1990 (over 13,000 treatment and 4000 control subjects), the Indiana Welfare Reform Evaluation in 1995 (over 67,000 treatment and 4000 control subjects), and the LA Jobs First GAIN evaluation in 1995 (over 15,000 treatment and 5000 control subjects).<sup>22</sup> Among these, only the GAIN evaluation discussed above allows inference about the role of JSA alone. The findings confirm that JSA can yield substantial gains in employment, at least in the short term.

In parallel, another group of experiments evaluated JSA services provided to recipients of UI. Most of these included a combination of direct JSA, instructions on how to search for a job, and verification of job search. These experiments, to a large extent discussed in [Meyer \(1995\)](#), included Nevada (1977, 1988), Charleston (1983), Texas (1984), New Jersey (1986), and Washington State (1986). Another set of experiments during same period assessed only the effect of verification of job search requirements. [Ashenfelter et al. \(2005\)](#) discuss experiments in Connecticut, Massachusetts, Tennessee, and Virginia.<sup>23</sup>

As summarized by [Meyer \(1995\)](#), a core finding of these studies is that JSA reduces UI receipt, at least in the short run. The effects are small, but cost-effective from the point

<sup>22</sup> There also have been evaluations of JSA services explicitly directed at low-income youth, but most such RCTs that we found were relatively small. The evidence on this subject quoted most frequently is related to the job search component provided in the JTPA and Jobs Corps programs.

<sup>23</sup> Other such experiments include Minnesota (1988), Maryland (1994), and Washington DC/Florida (1995–96), see [Greenberg and Schröder \(2004\)](#).

of view of the UI agency. The effects on earnings tend to be imprecise, consistent with the possibility that the program impacts derive from workers who leave the UI system without finding jobs. Little is known about which components of JSA matter. Experiments in Nevada and Minnesota suggest that intensive JSA has much stronger effects than do more limited treatments. There is mixed evidence as to whether the verification requirement alone matters: the experiments discussed in [Ashenfelter et al. \(2005\)](#) indicate no effects, while a Maryland study summarized in [Klepinger et al. \(2002\)](#) did. This question is a key aspect of ongoing evaluations of the Reemployment and Eligibility Assessment (REA) system, discussed below.

Since this early wave of UI experiments, the component of the UI system offering JSA and training has been repeatedly reformed, with several evaluations along the way. The Worker Profiling and Reemployment Services (WPRS) program was instituted in 1993. Under the WPRS, states are required to profile their UI claimants to identify those most likely to exhaust UI benefits and refer them to employment-related services.<sup>24</sup> This program was evaluated via a natural experiment in Kentucky beginning in 1994 ([Black et al., 2003](#); [Black et al., 2007](#)). The findings from the WPRS study suggest that receiving a letter asking individuals to come into the office for JSA services alone reduces UI receipt and raises earnings. An important open question is whether this influential finding is replicated in a true RCT and in less favorable labor market conditions.

The WIA of 1998 combined most job placement services and training services provided under the auspices of the federal government under one roof, the so-called one-stop centers (e.g., [Jacobson, 2009](#)). These centers, renamed America's Jobs Centers in 2012, provide both "core" employment services (e.g., JSA) and "intensive" WIA services (e.g., career counseling and training) to the three core constituencies—unemployed worker, welfare recipients, and hard-to-employ young workers.

As the structure of service provision has evolved, additional RCTs have evaluated the system's effectiveness at placing workers. For example, in 2005 the Department of Labor's Employment and Training Administration launched a program called REA, mandatory in-person visits aimed at speeding the reconnection of UI claimants to the workforce.<sup>25</sup> The REA meeting includes an eligibility review, provision of labor market

<sup>24</sup> The services include (1) an orientation session to explain what reemployment services are available; (2) an assessment of the claimant's specific needs; and (3) development of an individual plan for services based on the assessment. Claimants referred to reemployment services must participate in them as a condition of continuing eligibility. Allowable services include job search assistance and job placement services, such as counseling, testing, and providing occupational and labor market information; job search workshops; job clubs and referrals to employers; and other similar services.

<sup>25</sup> The REA program was instituted to counteract the trend toward processing of UI claims by telephone and the Internet. The concern was that the net effect of these changes was to reduce in-person contact and hence the opportunity to monitor job search activity and orient UI claimants to services available to speed their reemployment (e.g., [O'Leary, 2006](#)).

information, development of a reemployment plan, and referral to more specific reemployment services. The first wave of randomized evaluation of the effectiveness of the REA counseling process took place in nine states beginning in 2005; a second wave of evaluations took place in four states in 2009. In both cases, the evaluations found that the REA requirement and services reduce UI benefit receipt (Benus et al., 2008; Poe-Yamagata et al., 2011). Earnings outcomes were studied in only one state (Florida) and were positive. An ongoing REA evaluation examines the difference in the effect of enforcing the interview requirement alone relative to the combined effect of the interview plus services (Klerman et al., 2013). A simultaneous evaluation begun in 2011, the WIA Adult and Dislocated Worker Programs Gold Standard Evaluation<sup>26</sup> complements the evaluations of REA, WPRS, and earlier JSA programs by focusing on the effectiveness of WIA's intensive *and* training services geared to unemployed adults not covered by the earlier evaluations.

Summarizing the wide range of studies of JSA indicates important heterogeneity of effects by the population targeted. For welfare recipients, a difficulty in assessing the effect of JSA is that many experiments tested JSA in conjunction with other programs. Those studies that focus mainly on the effects of JSA, such as the randomized evaluations of WIN, NEWWS, or GAIN, often find positive effects on employment and earnings and negative effects on welfare receipt (but mixed effects at best on total income). These effects tend to be short-run, and less is known about the longer-term outcomes. There is also little known about the potentially important role played by context, such as local labor market conditions.

In studies of JSA for UI recipients, a common result is a precisely estimated but rather small effect—e.g., a reduction of about 1 week of UI benefits, with no corresponding positive effect on earnings—unless the services provided are very intensive. The frontier in this area is assessing to what extent these effects arise from the threat of enforcement of service requirements spelled out by law, basic JSA themselves, or more intensive services.

### 3.4 Practical aspects of implementing social experiments

Clearly, the implementation of large-scale social experiments is complex and faces a range of practical hurdles that can affect the quality of the results. Sections 2.3 and 4 of this chapter focus on a number of design issues that can limit the ability of even an ideal experiment to provide answers to the questions of interest.

Beyond these conceptual design issues, there are some common challenges and practical considerations that have come up over and over in the conduct of social experiments in the labor market. These play important roles in influencing the topics and questions that are studied via social experiments and in informing the study designs.

<sup>26</sup> See <http://www.mathematica-mpr.com/our-publications-and-findings/projects/wia-gold-standard-evaluation>.

One set of challenges derives from the fact that, as noted above, one of the defining characteristics of social experiments is that they intend to examine programs that are already in place or might be put in place in essentially the same form that was used in the experiment. For this purpose, the experimental samples and hence the sampling frame need to be representative of the population that the program serves. This is a challenge in the case of many labor market programs, in part because the sampling frame is often available only to program operators or the government, and may be difficult to access due to formal approval processes.

Once the sampling frame is obtained, it is necessary to randomly assign some members of the sample to the program of interest and others to a control condition, which might be exclusion from the program or an alternative program design. This, too, can be difficult when the program is already in place. For example, if the program in question exists within an ecosystem of other programs, services, and service providers, it may be hard to exclude participants from the program or, if this is done, to avoid also excluding them from other programs that are administratively integrated. For example, excluding a participant from JSA offered under the WIA might also in practice exclude him or her from job training and other programs, as the same offices that provide JSA also do screening and referrals for other services. While some of these problems might be reduced by studying programs *not* already in place, as in the case of the Negative Income Tax experiments or the NSWD, this can be quite costly, as the sorts of programs typically studied involve substantial program costs—commonly in the thousands of dollars per participant.

A second group of challenges has to do with the difficulty of enforcing compliance with randomization after it is conducted. Again, the use of actual programs tested in real-world settings limits the options. A common challenge in early experiments was that service delivery was delegated to individual case workers or sites that were both widely dispersed and not closely involved with the experimental design. This raises the possibility that caseworkers may deviate from random assignment, for example, ensuring that a potential participant viewed as especially needy is not assigned to the control group. For example, a key concern in the National Job Corps Study was to ensure that local program operators properly implemented the randomization. Modern practice centralizes the random assignment process, carefully tracking participants' initial assignments to ensure that participants assigned to undesirable treatment conditions do not reenter the randomization to obtain a better assignment.<sup>27</sup>

A third set of challenges has to do with the measurement of participant outcomes. Once again, this challenge derives, in large part, from the use of real-world populations as experimental subjects and from the large and heterogeneous subject pools common in

<sup>27</sup> For a discussion of approaches to address this problem, including related software, see, e.g., Crépon et al. (2013).

social experiments. These make it more expensive to ensure high response rates than in smaller and more targeted field experiments.

In many cases this challenge can be addressed by using administrative data to measure some outcomes. Administrative records may come either from the program under study—for example, UI payment records for studies of job search incentives for UI recipients—or from other records from other government programs (e.g., tax records). While this can resolve the attrition problem at low cost, it is often contingent on government cooperation or approval. Such cooperation is more likely in large-scale social experimental evaluations of existing programs than in other types of studies. Administrative data can also limit the set of impacts that can be studied, potentially creating important ambiguities in the interpretation of estimated treatment effects. In the UI case, for example, it is not clear whether a negative effect of increased job search enforcement on unemployment benefit payments indicates that people are finding jobs faster, or just that many people are leaving the program before finding jobs as a way of avoiding onerous enforcement procedures.

#### **4. GOING BEYOND TREATMENT–CONTROL COMPARISONS TO RESOLVE ADDITIONAL DESIGN ISSUES**

Whether one is interested in structural parameters or program evaluation, many questions of policy or scientific interest in labor and public economics require going beyond the basic RCT design described in [Section 2.1](#). We discussed a number of these questions in [Section 2.3](#). Here, we discuss ways to extend the basic RCT design to provide answers to these questions.

We organize our discussion around the major potential design issues we mentioned in [Section 2.3](#). For each, we discuss proposed solutions and, where relevant, point out potential extensions and limitations. We begin by discussing studies that address aspects relating to *internal validity*, including SUTVA violations (e.g., potential general equilibrium effects) and endogenously observed outcomes. We then discuss studies that address *external validity* concerns, including site and subgroup effects; effects on subpopulations other than experimental compliers; hidden or multiple treatments; mechanisms for treatment effects; and studies of optimal or simply alternative policies.

In some cases, the identified issues can be addressed ex post (after an experiment is complete), generally by imposing additional structure. In many of these examples the additional structure imposed is justified by appeal to theoretical considerations and is just sufficient to extend the RCT to address a specific question and the design issue it raises. In that sense, the studies can be viewed as an effort to bridge pure experimental or quasiexperimental approaches, credibly identifying a limited number of (potentially composite) causal parameters, with more traditional structural estimation that obtains a fuller characterization of the economic problem via the imposition of substantial

additional assumptions. In the ideal case, they maintain the best of both worlds, although they also share some of the limitations of each.

Another possibility is to build the structural questions of interest into the design of the experiment *ex ante*. This can provide credible identification with even fewer structural assumptions than are required for after-the-fact analyses, although can sometimes require a quite complex—and potentially difficult to administer—experimental design. There are fewer existing examples of this, but we discuss them where appropriate.

We discuss each of the design issues identified earlier in turn. Our discussion is meant to highlight the different approaches, as well as to clarify the scope, potential, and difficulties that arise when extending inference from standard RCTs to a broader range of questions.

## 4.1 Spillover effects and stable unit treatment value assumption

Social experiments in labor economics typically occur in the context of the local or regional labor market. If the number of workers participating in the program is large relative to the relevant segment of the labor market, the program could have an effect on the labor market outcomes of the control group. This would be a violation of SUTVA—the difference in outcomes between treated and control individuals would differ from the overall effect of the program on the entire population relative to not implementing the program, which is often the effect of primary interest.

Many social experiments in the United States have not raised serious spillover issues, as the treated populations have been small relative to the local labor market. However, this may not be true for large experiments, such as the National Jobs Corps Study. Welfare experiments may also create spillover effects if labor markets for former welfare recipients are sufficiently segmented.

A related issue is that comprehensive program evaluations in many cases *should* include spillover effects that are not captured by small-scale pilot studies. If the pilot programs are eventually scaled to broader populations of low-income workers—which has happened, among others, in the case of welfare reform, of training provided through WIA, or JSA services provided by WPRS or REA—then the potential extent of spillover effects would nevertheless matter, since any spillover effect would have to be included in a welfare assessment of the program. This would create systematic differences between the outcomes of the pilot study and the program effects of interest.

### 4.1.1 Addressing the issue *ex post*

Despite its potential prevalence in social experiments in the labor market, relatively few studies have dealt directly with the issue of spillovers or other failures of SUTVA. A handful of studies have tried to estimate spillover effects directly using interregional comparisons (e.g., Blundell et al., 2004; Ferracci et al., 2010; Gautier et al., 2012). There are roughly two approaches, neither of which is able to fully identify the spillover effect. One

approach is to compare control group outcomes to those of observably similar individuals in areas where no one is treated. Of course, there may be other explanations for differences seen in this observational comparison. Another approach is to compare the effect of treatment across sites with different treatment intensity or labor market conditions. This is again typically an observational comparison, as in most cases neither the treatment site nor the size of the treatment group (and hence the amount of potential spillover) is randomly assigned. For example, [Hotz \(1992\)](#) discusses the nonrandom selection of sites for the JTPA evaluations. [Alcott \(2015\)](#) studies the sources of observed bias from site selection in a large electricity conservation experiment. A recent paper by Crépon et al. (2013; see also [Baird et al., 2015](#)), discussed further below, resolves this problem in the context of a JSA program by randomly assigning both the treatment and the number of workers treated.

Absent such a multistage experimental design, relatively few options are available to researchers to assess the degree of the actual or potential spillover effects present in the context of their evaluation. An area of research where spillover effects have received substantial recent attention is the analysis of the employment and welfare impacts of extensions in UI benefits. Here, spillover effects arise because treated and untreated individuals compete for the same positions; the degree of the spillover effect therefore depends on the job creation response to the treated group's labor supply change. To assess the potential degree of spillovers, one can in principle use estimates of the matching function to adjust microeconometric estimates of the effect of policy-induced changes in UI durations on unemployment duration or exit hazards for the presence of crowding.<sup>28</sup> Such ad hoc simulations are partial equilibrium in nature and could be interpreted as a short-run effect, when vacancies have not yet adjusted. [Landais et al. \(2015\)](#) specify a general equilibrium model of the labor market that incorporates both crowding and vacancy responses. In a standard, competitive search-matching model, the vacancy response to changes in labor supply is sufficiently strong to offset the crowding effect completely.

In the spirit of using random variation in the treatment across localities to assess the presence of spillover effects, a couple of recent papers have tried to exploit region-specific changes in policy-induced UI variation in the United States to assess the full effect of the policy on the entire labor market ([Hagedorn et al., 2015a](#); [Hagedorn et al., 2015b](#)). Since UI variations usually depend on economic conditions at the state level, these studies use border communities unaffected by the policy change as counterfactuals.<sup>29</sup> A concern

<sup>28</sup> One added difficulty in the case of UI is that in most cases in the United States the policy-induced changes in the level or duration of UI benefits are a function of labor market conditions—making it crucial to properly control the direct effect of local labor market conditions.

<sup>29</sup> A key practical difficulty there is that measures of unemployment rates at the substate level are often very noisy. Estimates using administrative employment data based on the universe of private employees show little sign of spillover effects ([Johnston and Mas, 2015](#)).

with this approach is that the presence of spatial spillovers between adjacent or related labor market areas would again constitute a failure of SUTVA.<sup>30</sup>

Another source of SUTVA failures are interactions between treatment and control participants. Such “dilution” effects can lead to an underestimation of the treatment effect. If possible, a typical approach to circumvent such interactions is to raise the level of randomization (say, from a subgroup within a site to a whole site). This approach can help to avoid interactions between individuals in the treatment and control groups. It does not resolve potential interactions between treated participants. This may be part of the mechanism of the treatment; it may also be a potentially unintended source of variation in treatment intensity that we discuss under site effects. In either case, when designing an evaluation, it would be valuable to consider ways of keeping track of social interactions, perhaps by asking about friends in a baseline survey or monitoring (or manipulating) the use of certain kinds of social media. Another valuable target for data collection are factors relating to how treatment was obtained or take up was decided. Such information may be used to stratify the analysis by the predicted degree of SUTVA violations or at least assess the potential for significant departures from SUTVA.

#### **4.1.2 Addressing the issue *ex ante* through the design of the experiment**

In some circumstances it may be possible to avoid, or study, spillover effects by appropriately structuring a randomized experiment. For example, in the spirit of the nonexperimental studies cited above, treatment and control groups could be chosen to be sufficiently distant to avoid spillover effects. Alternatively, the treatment group could be chosen to be sufficiently small that spillover effects are unlikely to be a problem. If the spillover effects themselves are of direct interest, the experimental manipulation could be combined with preexisting variation in the strength of potential spillover effects (e.g., across submarkets), if available. The risk of such ad hoc or hybrid approaches is to potentially lose comparability of the control group or to confound spillover with other variation in treatment effects.

A preferable approach if spillover effects are potentially present is to manipulate both the treatment and the size of the treatment group (and hence the amount of spillover) experimentally. [Baird et al. \(2015\)](#) develop this strategy formally. [Crépon et al., 2013](#) implement it in the context of a public program assisting unemployed workers in their search for a job in France. The researchers manipulate both who gets assigned into the JSA program *within* a region (the classic experimental design), as well as randomly vary *between* regions the share of individuals assigned to the treatment group. The manipulation of both regional treatment share and individual treatment status allows separate

<sup>30</sup> [Cerqua and Pellegrini \(2014\)](#) develop alternative estimates to the TOT that take into account the degree of spatial spillover effects. Hagedorn et al.’s papers have been quite controversial; see, for example, responses from [Chodorow-Reich and Karabarounis \(2016\)](#) and [Coglianese \(2015\)](#).

experimental identification of the effect of the program holding the spillover effect constant and the combined program and spillover effects at various treatment intensities. The latter parameters are ultimately relevant for a cost–benefit or welfare analysis of the program and for extrapolation to alternative policy settings.

Similar strategies are available for other SUTVA failures, arising, for example, if some individuals in the control group get accidentally treated, or if treatment compliance depends on the take-up rate among peers. In some cases, one may choose the experimental setting to try to minimize SUTVA problems. For example, one can devise strategies to limit the potential for noncompliance (e.g., in case of web-based information treatments, access could be based on hardware address rather than passwords).

Another potentially interesting strategy is to make the degree and structure of SUTVA violations part of the analysis, as in the discussion of spillovers above. This may provide insights into the “black box” of how a program might work in a real-life setting and hence enhance external validity.<sup>31</sup> For example, one could experimentally vary the number of treated units in a reference group or network (e.g., classrooms, friends, etc.), examining interactions among individual treatment status, group treatment share, and perhaps also predetermined factors (such as the tightness of the group) that determine the degree of departure from SUTVA. Depending on the context, it may be possible to more explicitly manipulate interactions between individuals by introducing an additional treatment to the experimental design—for example, a forum in which interactions are facilitated.

## 4.2 Endogenously observed outcomes

In many labor market experiments, key outcomes include measures observed only for individuals who are employed, such as hours worked and wages. Hence, the impact of, say, welfare-to-work programs or job training programs can only partially be assessed based on simple RCTs alone. Although many studies report experimental impacts on the endogenously observed outcomes, these are understood to suffer from serious selection problems. In the same way, nonrandom attrition in follow-up data collection can bias the results of nearly any evaluation.

To illustrate, consider a program aimed at unemployed workers that includes skill development and JSA modules. We are interested in whether the program raises the probability that a participant is employed 1 year after participation and whether it makes them more productive when employed. For simplicity, we assume that participation is randomly assigned and compliance is perfect.

<sup>31</sup> Note that there is a parallel here with the issue of treatment compliance and heterogeneous treatment effects. Here, the compliance function is assumed to depend on treatment status of other individuals, and hence experimentally manipulating compliance probabilities is presumably more complex. Yet, as in the standard case of heterogeneous treatment effects, for external validity it is important to trace out the potential compliance-related interactions as fully as possible.

We have two outcomes here. We denote employment status by  $y_i = D_i y_{1i} + (1 - D_i) y_{0i}$ . For those who are employed at the follow-up survey, we observe the wage  $w_i = D_i w_{1i} + (1 - D_i) w_{0i}$ . Treatment effects of the program on the two outcomes are  $\tau_i^y$  and  $\tau_i^w$ . (We can imagine that  $w_{di}$  is well defined for an individual with  $y_{di} = 0$ ,  $d = \{0, 1\}$ , but simply not observed. It can be thought of as the individual's *latent* productivity that which he/she would be paid if a job were found.)

Estimation of  $E[\tau_i^y]$  is straightforward, as discussed above. But the impact on wages is much harder. In general, it is not possible to identify the ATE  $E[\tau_i^w]$ ; the TOT effect  $E[\tau_i^w | D_i = 1]$ ; or even the ATE for the subpopulation that would have been employed with or without the program (for whom  $\tau_i^w$  is least problematic),  $E[\tau_i^w | y_{0i} = y_{1i} = 1]$ .

The problem here is that it is impossible to distinguish, within each  $D_i$  group, between those workers who would also have worked in the counterfactual and those who would not have. Consider the treatment-control difference in mean observed wages:

$$\begin{aligned} & E[w_i | y_{1i} = 1, D_i = 1] - E[w_i | y_{0i} = 1, D_i = 0] \\ &= E[w_{0i} + \tau_i^w | y_{1i} = 1, D_i = 1] - E[w_{0i} | y_{0i} = 1, D_i = 0] \\ &= E[\tau_i^w | y_{1i} = 1, D_i = 1] + (E[w_{0i} | y_{1i} = 1, D_i = 1] - E[w_{0i} | y_{0i} = 1, D_i = 0]) \\ &= E[\tau_i^w | y_{1i} = 1, D_i = 1] \\ &\quad + (E[w_{0i} | y_{0i} = 1, y_{1i} = 1, D_i = 1] - E[w_{0i} | y_{0i} = 1, y_{1i} = 1, D_i = 0]) \\ &\quad + (E[w_{0i} | y_{1i} = 1, D_i = 1] - E[w_{0i} | y_{0i} = 1, y_{1i} = 1, D_i = 1]) \\ &\quad - (E[w_{0i} | y_{0i} = 1, D_i = 0] - E[w_{0i} | y_{0i} = 1, y_{1i} = 1, D_i = 0]). \end{aligned}$$

The first term here is the ATE in the subpopulation that works under treatment. It may not equal the overall ATE, but insofar as the potential wages of those who do not work are not relevant to social welfare, it is arguably the parameter of interest. The second term solely reflects selection into treatment and is zero under random assignment. But the third and fourth terms have to do with selection into employment, not selection into treatment. Random assignment does not ensure that they are zero, and the treatment-control contrast among workers may therefore be badly biased relative to the impact on wages for any fixed group of workers.<sup>32</sup>

<sup>32</sup> Consider a training and job search assistance program. Suppose 60% of workers will be always low productivity ( $w_{1i} = w_{0i} = w^L$ ), 20% will be always high productivity ( $w_{1i} = w_{0i} = w^H$ ), and 20% will become high productivity if exposed to the training sequence ( $w_{0i} = w^L$ ,  $w_{1i} = w^H$ ). All of the second and third groups will find jobs, with or without search assistance ( $y_{0i} = y_{1i} = 1$ ), but those in the first group of low-skill, impossible-to-train workers will find work if and only if they receive search assistance ( $y_{0i} = 0$ ,  $y_{1i} = 1$ ). In this setting, the program's average treatment effect on employment is 0.6; the average effect on latent productivity is  $0.2*(w^H - w^L)$ ; and the average effect on wages of those who would work with or without the program is  $0.5*(w^H - w^L)$ . The estimated treatment effect on wages conditional on employment is  $-0.1*(w^H - w^L) < 0$ . Selection has led to a perverse estimate here: the training program has a positive effect on 20% of participants and a negative effect for no one, but the experiment appears to indicate that it reduces earnings.

One fallback approach is to examine only the program's effect on the share of participants earning high wages, treating low-wage workers and nonworkers the same. This effect can be estimated without bias. Another fallback is to include the nonemployed in the wage analysis, with wages set to zero. This in some cases is the impact of interest in any case and is correctly identified by the experiment. However, it is quite misleading if interpreted as the magnitude of the effect on productivity, either for the full population or for the subgroup that would have been employed with or without treatment. Without an ability to measure *counterfactual* employment status at the individual level, the latter effects are not identified.

#### **4.2.1 Addressing the issue ex post**

Nonrandom attrition in particular has been a long-standing concern in the experimental literature in labor economics (e.g., [Hausman and Wise, 1979](#)). A classic experimental design would be deemed successful if attrition is low and balanced in terms of magnitude and observable characteristics between the treatment and control groups. If this is the case, reweighting the samples may still recover the effect of the TOT or LATE among the original set of compliers (e.g., [Ham et al., 2011](#)). Yet, there are relatively few explicit attempts in the literature to address selection bias in other contexts.

A large literature in labor economics has dealt with sample selection problems, especially in the analysis of wages and hours in the context of the classic human capital and labor supply models. Largely based on that literature, here we will review several approaches to deal with selection bias: the use of control functions to address selection; estimation of percentile effects instead of mean impacts; use of additional data to control for selection; construction of bounds based on selection probabilities; and construction of bounds using theory.

##### **4.2.1.1 Parametric selection corrections**

The “classic” approach to control for selection bias in estimating the effects of treatment effects on wages or hours worked is based on control functions. Labor supply theory, along with parametric assumptions, is used to derive an explicit expression for the selection bias in terms of the participation probability, which under monotonicity determines the amount of sample selection. This is then accounted for directly in the outcome equation (e.g., [Gronau, 1973](#); [Heckman, 1979](#)).

Early on it was recognized that absent experimental variation in participation (e.g., an exogenous instrument affecting only participation and not the outcome equation), identification is only based on functional form assumptions, and results can be quite misleading if these assumptions are even slightly incorrect. By contrast, a substantial literature has shown that once an instrument for participation is available, treatment effects in the outcome equation can be identified under quite general functional form and distributional assumptions (e.g., [Newey et al., 1990](#)). For example, [Ahn and Powell \(1993\)](#)

show that under assumptions of a single, strictly monotonic index for selection, variation in the probability of participation independent from the variables in the outcome equation suffices to control for selection. The difficulty is, of course, that often such independent source of variation is not available.

[Card and Hyslop \(2005\)](#) consider a special case in which an RCT does generate exogenous variation in participation: an employment subsidy program. They show that if the program only has positive effects on labor supply and does not affect the wages for those who would have worked without it, then the experimental effect on the hourly wage can be consistently estimated by the ratio of the treatment effect on total earnings divided by the treatment on total hours worked.

Card and Hyslop's assumptions are inappropriate for any program designed to affect wages and not just participation. Below we discuss how the experimental design itself may be modified to obtain exogenous variation in participation, even in programs with effects on multiple margins.

#### 4.2.1.2 Non- and semiparametric selection corrections

Absent an instrument for participation, in the presence of selection the treatment effect on mean wages is not identified. However, several studies have exploited the fact that under certain assumptions quantile treatment effects (QTEs) may be consistently estimated even in the presence of selection. A QTE for the  $q$ th quantile is defined as the difference in the  $q$ th quantile of the outcome distribution in the treatment and control groups, respectively.<sup>33</sup> It is not necessary to observe each individual's outcome to compute the  $q$ th quantile; it suffices to know that someone is above or below that quantile. Thus, if one can assume that all those who are not employed have potential wages in the bottom  $q$  percent of the distribution, one can estimate the treatment effect on the  $q$ th quantile of potential wages by merely assigning all nonworkers the minimum observed value (e.g., [Powell, 1984](#); [Buchinsky, 1994](#)). Hence, under this assumption all quantiles above the value of the rate of nonemployment of the respective group can be identified. The lower value of nonemployment of the treatment and control group determines which QTE can be identified.

A variant of this approach is to examine the simple treatment-control difference in the probability of being observed in employment with a wage greater than some relatively high threshold  $w^*$ . For many program evaluations, understanding the impact on this outcome may be sufficient—it may not matter greatly whether the impact derives from moving some people from nonemployment into high-wage employment or from simply lifting those who would have worked anyway into higher-wage jobs.

<sup>33</sup> For any random variable  $Y$  having cumulative density function  $F(y) = \Pr[Y < y]$ , the  $q$ th quantile of  $F$  is defined as the smallest value, such that  $F(y_q) = q$ . If we consider two distributions  $F_0$  and  $F_1$ , then  $\text{QTE}(q) = y(1) - y_q(0)$ , where  $y_q(g)$  is the  $q$ th quantile of distribution  $F_g$ .

And even when the latter component is the one of interest, this would be identified so long as those pulled into employment by the treatment have wages that are uniformly below  $w^*$ .

It is not clear, however, that the required assumption holds—as pointed out by [Altonji and Blank \(1999\)](#), among others, at any given time, some high-wage individuals may be nonemployed. Moreover, this strategy is only useful insofar as differences in quantiles of the outcome are deemed sufficient for evaluating the effect of the program.

Another approach uses reservation wages to measure selection into the subsample of observed wages. This works because—if correctly measured—the reservation wage captures the lowest wage for which an individual is willing to work. Hence, the reservation wage provides the censoring point for an individual's wage—offer distribution, allowing one to make inferences about potential wages for those individuals not working in the treatment and control group. [Johnson et al. \(2000\)](#) use the minimum of all observed wages for an individual in longitudinal data to bound the reservation wage, under the assumption that it is stable over time. [Groger \(2005\)](#) uses directly reported reservation wage information from a randomized evaluation of Florida's Family Transition Program, a welfare-to-work program with emphasis on work incentives and time limits. With this information, he estimates the treatment effect of the program on wages using a bivariate, censored regression model that allows for classical measurement error in both observed wages and reservation wages. Once [Groger \(2005\)](#) controls for selection, he finds the program had statistically significantly positive effects on wages.

Addressing the selection problem using direct measures of reservation wages makes intuitive use of the reservation wage concept. Moreover, often information on reservation wages is already being collected in the context of programs providing JSA, or if not they are at least in principle relatively easy to elicit if the experimental design includes a survey component. However, recent research suggests that in practice reported reservation wages appear to only partly reflect the properties of the theoretical concept (e.g., [Krueger and Mueller, 2016](#)), casting some doubt on the robustness of this approach. In particular, Krueger and Mueller report that a substantial number of workers accept (reject) jobs offering wages below (above) their reservation wage, implying that care should be taken in using reservation wages of the nonemployed to make inferences about unobserved wage offers.

Yet another approach is to attempt to derive bounds for the treatment effect under conditions more general than the monotonicity assumption inherent in the [Ahn and Powell \(1993\)](#) and similar estimators. This allows researchers to investigate how severe the bias from selection could possibly be and what can be learned under general assumptions rather than to try and to obtain a point estimate under more restrictive assumptions.

One bounding approach is proposed by [Horowitz and Manski \(2000\)](#). This strategy asks how much the estimated treatment effect would be inflated if all missing treatment observations were assumed to have the highest possible outcomes and all missing control

observations the lowest; then it asks how much it would be depressed if the opposite assumptions were made. Unfortunately, these bounds are typically not very tight, particularly when the outcome variable's support is potentially unbounded as, for example, in the case of wages.

[Lee \(2009\)](#) proposes a strategy for obtaining tighter bounds, via stronger assumptions: he assumes that anyone not employed in the control group would also have been nonemployed had they been in the treatment group, so that selection bias arises solely from participants in the treatment group who are employed but would not have been had they been assigned to be controls.<sup>34</sup> He can then bound the treatment effect by making extreme assumptions about this latter group. Denote the excess fraction employed in treatment group by  $p$ . The upper (lower) bound is constructed by removing the lowest (highest) fraction  $p$  observations from the treated subsample and recomputing the mean outcome for the treatment group—effectively making the worst-case assumption that selection was fully responsible for the entire upper or lower tail of values. [Lee \(2009\)](#) shows that the resulting bounds are sharp and provides formulas for the standard errors. In the case of Job Corps, the procedure results in informative bounds suggesting positive wage effects from training—albeit a zero effect is contained in the confidence interval.

[Lee's \(2009\)](#) approach based on trimming requires relatively weak assumptions. It presumes only that selection is monotonic in the treatment—that treatment either only increases, or only reduces, selection into employment. Monotonicity is implied by standard empirical binary choice models typically used to model participation choices (e.g., [Vytlačil, 2002](#)), and hence bounds based on trimming are applicable to a wide range of problems, including selective employment, survey nonresponses, or sample attrition.

If one is willing to impose further structure from theory, one may obtain tighter bounds more specific to a particular problem. This is especially useful if the theory has explicit predictions about how the endogenous outcome responds to incentives.<sup>35</sup> This is pursued by [Kline and Tartari \(2016\)](#), who analyze the randomized evaluation of Connecticut's Jobs First welfare-to-work program. While previous analyses had found only small responses in hours (the intensive margin), absent an instrument for participation (the extensive margin), sample selection makes such estimates hard to interpret. [Kline and Tartari \(2016\)](#) use revealed preference arguments in the context of a canonical but nonparametric static labor supply model to describe which observed responses to the treatment at the intensive and extensive margins are consistent with the theory. Given the

<sup>34</sup> The role of treatment and control groups are reversed if the treatment reduces employment.

<sup>35</sup> This may be more easily done for hours, which is typically assumed to be a choice variable, than for wages. Yet, to some degree, wage may be a choice variable as well, for example, if jobs offer wage and effort combinations among which workers choose. This is the approach taken in some modern public finance, which often substitutes hours worked with taxable earnings as the choice variable in analyses of intensive-margin labor supply.

nature of the program studied, the result is a mapping of discrete counterfactual outcomes (including nonparticipation as well as participation at different intensities) under treatment and nontreatment, with restrictions on the allowable counterfactuals. The question then is how likely certain transitions are, and in particular whether changes at the intensive and extensive margins occur with positive probabilities. Since Kline and Tartari can only observe the marginal distribution across states for the treatment and control groups, they cannot point-identify the transition probabilities. Instead, they construct bounds for transition probabilities among the entire (discretized) distribution of states, including the probability of changes in the intensive margin due to the treatment. Their approach also allows them to test the restrictions from the model.

This approach is useful, since it allows [Kline and Tartari \(2016\)](#) to learn about intensive margin responses to the Jobs First program in the presence of selection. Their results could also be used to think about the likelihood of intensive margin responses for similar programs in similar populations. Alternatively, the estimated bounds from the matrix of transition probabilities could be used, along with the marginal distribution of labor supply under an existing program (AFDC, the program of the control group), to construct bounds for the intensive and extensive labor supply responses that could arise if Jobs First was implemented at another site. A potential issue is that the procedure is complex and the analysis is specific to the Jobs First program. Hence, while the general approach may be applicable to a range of problems, this would require careful specification of the decision problem, of the restrictions imposed by revealed preference theory, and of counterfactuals for each case. Nevertheless, since many social experiments are concerned with welfare and other programs that provide explicit variation in employment incentives and hence useful information on the likelihood of counterfactual outcomes, it is useful to consider the role that theory can play in providing bounds on treatment effects on endogenous outcomes.<sup>36</sup>

#### **4.2.2 Addressing the issue *ex ante* through the design of the experiment**

The endogenous outcome problem is often easily anticipated when designing an experiment, as it arises whenever outcomes such as wages or hours are of interest and nonemployment is a realistic possibility. There are various ways to adjust the experimental design to facilitate analysis of potential sample selection bias. For example, suppose in the case of the effect of a training program on wages, the researcher believes that there are exogenous factors determining a worker's labor supply decision. If these factors can be measured *ex ante*, the randomization could be stratified by the likelihood of employment as predicted by the exogenous instruments. Stratification would ensure sufficient sample sizes in each exogenous labor supply tier. (If only available *ex post*, say, in a

<sup>36</sup> Similar approaches have been pursued in [Blundell et al. \(2011\)](#).

follow-up survey, even absent stratification, such variables can be still used as instruments for participation if sample sizes are sufficiently large.)

However, as it is usually difficult to come by good instrumental variables, the real power of a well-designed RCT would be to manipulate sample selection directly. In the training example, this would entail adding a second source of randomization that explicitly modifies the incentive to work (or the likelihood of finding a job) but does not otherwise affect the endogenous outcome. Whether this is feasible depends on the context. However, sample size considerations need not be a hurdle to adding a second treatment, since with cross-classified treatments the addition of a second treatment has little effect on the power for analyzing the effects of the first in isolation. This approach is particularly useful if one is interested in external validity, since the two-dimensional experimental variation may allow one to trace out the treatment effect of training for subpopulations with different employment probabilities.

In the case of nonrandom attrition, a version of this approach would be to randomly select a group of participants to follow up more intensively, perhaps stratified within groups with different ex ante attrition probabilities. The contrast between mean outcomes in this subgroup and for other participants (again, perhaps within strata) identifies the selectivity of attrition and can be used to adjust the full-sample estimated treatment effects. This is the approach pursued in the follow-up waves of the Moving to Opportunity (MTO) experiment (e.g., [Kling et al., 2007](#)). Another solution worth pursuing is to obtain administrative data for the universe of initial participants, including those who have failed to respond to follow-up surveys. Although these data can also be selected—they typically do not include earnings from informal jobs—the selection is different from that created by survey attrition, so the combination of sources can be valuable (though sometimes confusing, as in the Job Corps evaluation discussed above). Since merges to administrative data can be usually only conducted only with identifying information from the survey and permission from participants, it is a good idea to factor the need for additional data into the initial research design.

### 4.3 Site and group effects

In many cases an essential problem is to identify the subpopulations that benefit most from a program, so as to target them for treatment. However, there are often many possible subgroups to examine. When many comparisons are estimated, the chance of a false discovery—a treatment—control contrast that is statistically significant, even though the true treatment effect is zero—rises toward one. Avoiding incorrect inferences in such a setting requires care.

A version of the subgroup effects problem is to identify variation in treatment effects across program locations or sites. Such variation might arise from observed local characteristics—e.g., treatment effects of training or job search experiments may depend

on the tightness of the local labor market. Where the relevant characteristics of the labor market are clear *ex ante* and their dimension is limited, this is relatively straightforward. But if the relevant dimensions are not clear or the number of potential contrasts is large, the multiple-comparisons problem becomes relevant. Alternatively, there might be unintended variation in treatment intensity or in the fidelity or effectiveness of treatment delivery among treatment sites. Such site effects render the interpretation of the estimated treatment effect of the overall treatment difficult and limit external validity. If they are potentially important, we need estimates of each site's separate effect. This implies that there are as many treatment effects to be estimated as there are sites at which the experiment is implemented.

A conceptual issue in evaluating the success of social experiments with site variation is to decide whether the parameter of interest is the effect of the program in its most successful variants, with strong local partners and appropriate local conditions, or the average effect across a range of local circumstances. When the latter is of interest, the ideal experimental design would involve drawing participants from all sites. But this is often impractical. More commonly, social experiments have been carried out at one or a few sites. These are often chosen because the local management is willing to participate, or because they are seen as exemplars of the program. This makes it difficult to interpret the experimental results as representative of the program as a whole (see, e.g., [Hotz, 1992](#); [Alcott, 2015](#)), but may come closer to identifying the program effect under close-to-ideal circumstances.<sup>37</sup>

### **4.3.1 Addressing the issue *ex post***

On its face, it is straightforward to estimate heterogeneity of treatment effects along observed dimensions (e.g., race, gender, or past work experience) using data from an already-completed randomized trial: one simply constructs treatment–control contrasts separately for each subgroup. Many authors emphasize the importance of conducting the randomization separately for each subgroup of interest. This is not in principle necessary—unconditional random assignment ensures that assignment is random conditional on predetermined characteristics as well—but can add power for subgroup comparisons, especially in smaller samples.

A more important issue is the potential number of comparisons to be estimated. If enough subgroup estimates are computed, even a program that has no effect on anyone will be likely to show a statistically significant effect for some subgroup. (A similar

<sup>37</sup> A related but distinct problem is the question of ensuring “fidelity of implementation” in an RCT—a close alignment between the program’s intended design and the services that are actually delivered. While this is important for maximizing the statistical power of the experiment and for testing whether the program’s theory of action is correct, it limits the external validity for use in making judgments about the likely overall impact of real-world programs, which may not be implemented with high fidelity.

problem arises when considering effects on multiple outcomes.) Researchers have taken a number of approaches to this multiple comparisons problem. One is to specify the subgroups that will be considered, and the hypotheses of interest, before analyzing the data. This can limit the scope for unconscious data mining. It also ensures that the number of comparisons that were considered is known, so that the *p*-values of simple treatment-control contrasts can be adjusted for the multiplicity of the comparisons being estimated. An appropriate adjustment makes it possible to obtain accurate *p*-values for the test of whether the program had any effect on any subgroup. But two issues remain: these tests typically have *very* low power. In addition, even when they do reject they are often not able to identify *which* subgroups have nonzero treatment effects. A full discussion of adjustment for multiple comparisons is beyond the scope of this chapter, but [Anderson \(2008\)](#) is a useful reference.

Multiple comparisons approaches can be useful as well for the analysis of treatment effects by site and/or provider. But the questions of interest regarding site effects are not generally whether each site's effect is or is not different from zero, which is what multiple comparisons adjustments are designed to answer, but rather the magnitude and correlates of variation in treatment effects across sites. Moreover, the fact that the site-specific treatment effects can in some sense be seen as draws from a larger distribution opens up new options for analysis that are not available in traditional studies of subgroup treatment effects.

The mid-1990s National Job Corps Study, discussed above, illustrates some of the issues involved.<sup>38</sup> As mentioned previously, the random assignment study indicated that the program has a positive average effect on earnings 4 years after participation, of a magnitude roughly comparable to the return to a full year of education ([Schochet et al., 2008](#)). (At the time of the evaluation, the average participant was enrolled for about 8 months.)

But like other job training programs, the specific “treatment” provided to Job Corps participants varies substantially across individuals, according to perceived needs. Moreover, Job Corps services are delivered at 110 mostly residential centers, the majority of which are operated by private contractors. Some providers may be better at delivering an effective program (or at guiding participants to the types of services that they need) than are others. The center-specific treatment effects are thus of great interest.

The Department of Labor (DOL) has long used a performance measurement system to track performance of the different centers and inform decisions about contract renewal. Performance measures are nonexperimental and include statistics such as the GED attainment rate or average full-time employment rate of program participants at

<sup>38</sup> Other studies that examine similar questions are [Bloom et al. \(2005\)](#) and [Barnow \(2000\)](#). See also our discussion of treatment spillovers above.

each center. But it is not clear that these performance indicators successfully distinguish center impacts from differences in the populations served by the various centers.

[Schochet and Burghardt \(2008\)](#); hereafter “SB”) attempt to use the random assignment Job Corps Study to validate DOL’s performance indicators (see also [Barnow, 2000](#); who carries out a similar exercise for JTPA). In principle, estimation of site-level causal effects using the experiment is straightforward: one simply compares mean outcomes of the treatment and control groups at each site, relying on the overall random assignment to ensure balance of each site-level comparison. But a few challenges arise.

First, in the Job Corps Study, randomization took place before applicants were assigned to centers. Thus, treated individuals are associated with centers, but control individuals are not. SB address this by using intake counselors’ assessments of the center that the applicant would most likely attend, collected prior to randomization. To ensure that treatment and control individuals are treated comparably, they use this prediction for both groups, even when it differs from the actual treatment assignment. Differences occurred for only 7% of treatment group enrollees, largely because participants tend to enroll in the closest center or in one that offers a particular vocational program.

Second, even a large RCT sample—the Job Corps Study included over 15,000 participants—can have very small sample sizes at the individual site level. Rather than estimate center-specific treatment effects, SB divide centers into three groups based on their nonexperimental performance measures and estimate mean treatment effects for each group. Interestingly, they find that mean program impacts do not differ significantly across groups, suggesting that the performance measurement system is not successfully identifying variation in centers’ causal impacts. A related exercise is carried out by [Bloom et al. \(2005\)](#), who first estimate statistically significant variation in treatment effects across 59 local offices that participated in three welfare-to-work experiments, then use a multi-level model to estimate the relationship between office characteristics—mostly having to do with the way that the treatment was implemented in each site, although they also include the local unemployment rate—and office-level treatment effects. In contrast to the Job Corps Study, they do find significant associations of the treatment effect with both their implementation measures and the local unemployment rate.

[Bloom et al., 2005](#) interest is in identifying which program features are most effective. It is important to emphasize, however, that the association between site-level characteristics  $X_j$  and the site-specific treatment effect  $\tau_j$  is observational, not experimental, and does not bear a strong causal interpretation. It is quite possible that what appears, for example, to be a strong association between the emphasis that sites place on quick job placement and the site-level treatment effect instead reflects a nonrandom distribution of this emphasis across sites that vary in other important ways.

Like the Job Corps Study, [Bloom et al. \(2005\)](#) do not investigate variation in site impacts conditional on  $X_j$ . In many settings, that variation might be of substantial interest. One might like, for example, to estimate effects of individual sites, or to ask which of a

number of available performance measures do the best job of predicting experimental impacts. The latter question is a natural one to ask regarding the Job Corps Study, but to our knowledge it has not been pursued with experimental data (though see Barnes et al., 2014 for a related investigation using nonexperimental data).

Much work on the estimation of site effects themselves comes out of efforts to measure hospital, school, or teacher performance (see, e.g., Jackson et al., 2014; Rothstein, 2010). These studies are program evaluations, treating each site or teacher as a distinct “program,” but cannot rely on random assignment to identify program effects. As in the Job Corps Study, there are many sites but samples are frequently small at the site level, so—even if selection biases are set aside—site-specific treatment effect estimates are quite noisy. One consequence is that actual treatment effects will typically be closer to the average than are estimated effects, even when the research design permits unbiased estimation of each effect. Thus, it is common in these literatures to “shrink” the estimated treatment effects toward the mean. The procedure goes by many different names—e.g., shrinkage, Empirical Bayes, regularization, partial pooling, multilevel modeling—but the basic idea is that the posterior estimate of a site’s effect equals a weighted average of the unbiased estimate of that site’s effect and the mean site effect, with weights that depend on the precision of the site estimate.

Let  $\tau_j$  represents the impact of the program at site  $j$ , and suppose that across sites,  $\tau_j \sim N(\bar{\tau}, \omega^2)$ . Suppose that we have a noisy but unbiased estimate of the site  $j$  effect:  $t_j | \tau_j \sim N(\tau_j, \sigma^2)$ . Then the former can be treated as a prior distribution for  $\tau_j$ . By Bayes’ Rule, the posterior mean of  $\tau_j$  given the observed estimate is

$$E[\tau_j | t_j] = \bar{\tau} + f(t_j - \bar{\tau}),$$

where

$$f = \omega^2 / (\omega^2 + \sigma^2)$$

is the reliability ratio of the site-specific effect estimate.<sup>39</sup>

When the treatment effect varies systematically with site-level covariates—characteristics either of the treatment or of the counterfactual—this can be used to improve precision. If the site effects are modeled as a function of site characteristics,  $\tau_j = X_j \beta + \nu_j$ , with  $\nu_j \sim N(0, \sigma_v^2)$ , then the noisy site-level estimate  $t_j$  should be shrunk toward the conditional mean rather than to the grand mean:

$$E[\tau_j | t_j, X_j] = X_j \beta + f'(t_j - X_j \beta),$$

<sup>39</sup> The posterior mean is also known as an Empirical Bayes estimate. It is an unbiased predictor of the true site-level treatment effect  $\tau_j$ , if the site-specific estimates  $t_i$  are unbiased estimates (Rothstein, 2016).

where  $f'$  is the conditional reliability ratio,  $f' = \omega^2 / (\omega^2 + \sigma_v^2)$ . This is sometimes known in the statistics literature as “partial pooling.”

One use of the shrinkage approach is by [Kane and Staiger \(2008\)](#), who use a random assignment experiment to validate nonexperimental estimates of teachers’ treatment effects on their students. They shrink the nonexperimental estimates, under the assumption that these estimates are valid, and ask whether the result is an unbiased predictor of a teacher’s treatment effects under random assignment.

Kane and Staiger focus on “value-added” scores, estimates of teachers’ effects on their students’ test scores from observational regressions, as the sole nonexperimental estimate. They fail to reject the hypothesis that these scores are unbiased predictors of the experimental effects, consistent with the view that they are unconfounded by student sorting. But the experiment has quite low power to distinguish alternative explanations, and [Rothstein \(2016\)](#) argues that the question remains unresolved.<sup>40</sup>

[Angrist et al. \(2015\)](#) explore the optimal combination of experimental estimates with potentially biased but more precise nonexperimental estimates to obtain minimum mean-squared-error predictions of schools’ treatment effects. A related question is whether nonexperimental measures of other parameters (e.g., classroom observations) can improve the prediction of experimental effects. If so, one might want to use a weighted average of the available measures, weighted to best predict the experimental treatment effect, for performance measurement purposes. To our knowledge, no study has attempted to estimate these weights in an experimental setting (though see [Mihaly et al., 2013](#); for a nonexperimental analysis).

### 4.3.2 Addressing the issue *ex ante* through the design of the experiment

Ultimately, small sample sizes have limited analysts’ ability to identify site- or group-level variation in treatment effects. But there may be ways to design experiments to better support these investigations. Most obviously, resources can be put into collecting data on variation in the quantity and types of treatments delivered, to support analyses (like that of [Schochet and Burghardt, 2008](#) or [Bloom et al., 2005](#)) of how site treatment effects vary with observable measures of site treatment variation. Large-scale program evaluations often include implementation analyses alongside randomized impact evaluations, and if these two portions were closely integrated, the results of the implementation study could be used to inform an analysis of site effects in the impact evaluation sample. Power can also be improved by conducting randomization within site-level strata and by minimizing noncompliance rates (and carefully measuring treatments actually received).

<sup>40</sup> For more on the topic of teacher value-added, see [Chetty et al. \(2014\)](#) and [Rothstein \(2016\)](#).

## 4.4 Treatment effect heterogeneity and external validity

The empirical literature on program evaluation has been increasingly aware of the importance of potential heterogeneity in treatment effects for interpreting estimates of program impacts and assessing their external validity. Many evaluation samples are drawn from specific populations—individuals in particular regions or cities, individuals entering a program in a certain way, or individuals thought suitable for a proposed alternative program. If treatment effects vary, generalizing from these samples to a broader population is hazardous. Another variant of the external validity problem arises when the compliance rate in the experimental sample differs from what would be expected outside the experiment, as the experimental LATE may not correspond to an appropriate complier population for the program evaluation of interest.

There are several potential sources of heterogeneity. In the previous section, we have discussed differences in characteristics of the environment (such as state of the labor market, including business cycle and industry or occupation structure, population density, or labor market discrimination) and differences in aspects of the program (such as unintended differences in the intensity of treatment, something we address under site effects). In this section, we focus on the case where treatment effects vary because of differences in characteristics at the individual level (such as preferences, abilities, health, beliefs, resources, family environment, or access to networks). Below and in [Section 4.6](#), we also discuss variation in treatment effects arising because of variation in structural aspects of the program, such as differences in work incentives.

### 4.4.1 Addressing the issue *ex post*

The literature is broadly in agreement on how to deal with heterogeneity in treatment effects by *observable* characteristics of study participants. As discussed in [Section 4.3](#), the experimental design implies that one can obtain consistent estimates of the treatment impact for each subgroup, subject to having sufficiently large sample sizes. One can then extrapolate the TOT and ATE to settings with other distributions of observable characteristics by constructing appropriately weighted averages of subgroup effects and corresponding standard errors. As a more common alternative, one can directly estimate TOT and ATE for another population by reweighting the original sample to match the distribution of observable characteristics of the target population (e.g., [DiNardo et al., 1996](#)). If multiple treatment sites are available, in principle a similar approach can be used to assess the effect of environmental characteristics, such as labor market conditions or industrial structure.

The case of heterogeneity by *unobserved* characteristics has presented greater challenges. Unfortunately, the individual-level treatment effect is generally not identified either by experimental nor nonexperimental methods. Even with perfect compliance, an experiment identifies only the ATE conditional on observed characteristics.

Some argue that ATEs are sufficient for most purposes, as we care only about the distributions of outcomes under alternative policies and not about the positions of particular individuals within those distributions. This is a controversial claim, however—in many contexts, a program that helped some individuals but hurt others by an equal amount, with zero average effect, would be judged worse than nothing.

Moreover, average effects may not be generalizable beyond the population (with perfect compliance, experimental participants, or with imperfect compliance, the subgroup of compliers) identified by an experiment. With heterogeneous treatment effects, neither the TOT nor the complier LATE may be relevant for other populations of interest. A key question then is how representative the experimental compliers are of the group of people that would be potentially affected by the program in question. In many cases the program compliers are likely to be similar to the population of interest, in which case the complier LATE is likely to approximate the relevant parameter. In other cases—for example, when compliance is likely to differ between the study and the program at scale—the estimated LATE from one program evaluation may be less useful.

[Heckman and Vytlacil \(2005\)](#) propose a conceptual framework to analyze heterogeneity in treatment effects that relies on the concept of the marginal treatment effect (MTE). If  $\tau_i$  denotes the individual treatment effect,  $X_i$  is a vector of observed individual characteristics, and  $v_i$  is the error in the equation determining take-up of treatment, then the MTE is defined as  $E[\tau_i | X_i = x, v_i = v]$ ; of interest is how this varies with  $v$ . This structure provides a framework for considering external validity. The traditional LATE obtained from analyses of experiments with noncompliance can be seen as the integral of the MTE over a particular range of  $v$ , but proposals to expand or roll back programs may implicate MTEs at other  $v$  values.

To move beyond the LATE, we require a multivalued instrument that can map out the full distribution of  $v$  (or, equivalently, the full range of  $\Pr(T = 1 | X)$ ). If such an instrument is available, the MTE can be obtained by a nonparametric regression of the outcome on the fitted probability of program participation resulting from the first stage equation.<sup>41</sup>

This is not possible in the case of a simple RCT. However, when the RCT implemented at multiple sites, and if one is willing to assume that heterogeneity of site effects is limited to compliance rates with no variation in effects on the outcome, one can examine the relationship between the site-specific compliance rate and the site-specific estimated

<sup>41</sup> Many other relevant parameters, including LATE and ATE, can be expressed as functions of the MTE. However, to estimate the ATE or the TOT, say, one needs to obtain the MTE for each value of  $X$  for the full range of complier probabilities, i.e., from 0 to 1. While in many cases this may be infeasible due to data limitations, if available this could be used to extrapolate the ATE or TOT for populations with different compliance rates and distribution of characteristics.

treatment effect (i.e., the site-specific LATE).<sup>42</sup> (Alternatively, one could directly regress the site-specific treatment effect on the estimated probability of take-up and obtain the MTE for different compliance rates.) This relationship could in principle be used to forecast the LATE at a potential alternative treatment site (possibly reweighting to adjust for differences in observable characteristics), given a forecast of the new site's compliance rate. More generally, this approach would allow inferring the effect of any intervention affecting the cost of compliance and hence the compliance rate itself.

At times it is useful to go further to estimating the full distribution of treatment effects. The above method will not accomplish this. Heckman et al. (1997) show that without additional assumptions, experimental data are essentially uninformative about the treatment effect distribution. Moreover, they demonstrate that quite strong assumptions on the dependence of counterfactual outcomes in the control and treatment states are needed to obtain plausible estimates of the distribution of the effect of training in the context of the National JTPA Study. Nevertheless, as mentioned at the outset, knowledge of the distribution of heterogeneous treatment effects is undoubtedly important in assessing the impact of a particular program (although it is less straightforward how such information can be used to address the issue of external validity if treatment effects vary purely with unobserved characteristics).

One approach that has been used to make inferences about heterogeneity in treatment effects is estimation of QTEs. As discussed in Section 4.1, the QTE for the  $q$ th quantile is defined as the difference in the  $q$ th quantile of the outcome distribution in the treatment and control groups, respectively. It is clear that absent strong assumptions, such as rank stability, QTEs do not recover the distribution of treatment effects (although they do recover the effect of the treatment on the outcome distribution, which may be sufficient for many purposes; see Athey and Imbens, 2017, this volume). Yet, it can be a helpful and easy-to-implement diagnostic device in at least two senses. First, a QTE analysis can be used to test the assumption of constant treatment effects, which would imply that the QTE is equal at all quantiles. Second, in some cases particular features of a program allow one to derive predictions as to responses in different quantiles of the outcome distribution (see below). More generally, QTE may provide a broad descriptive sense of potential treatment responses.

One source of treatment effect heterogeneity is differences in the structure of the program to be evaluated. In this case, theory may provide weak assumptions that allow making inferences about the distribution of treatment effects. Welfare programs represent a good example, since they usually combine a range of different labor supply incentives

<sup>42</sup> Note that the weighting function of the LATE estimator for multivalued instruments in Angrist and Imbens (1995) is proportional to the differences in take-up probabilities between different values of the instrument (ordered by the values' impact on take-up). This difference can be interpreted as the difference in compliance between instrument values.

arising among others from welfare payments, earnings disregards, implicit tax rates, or phase-out regions. Clearly, these incentives interact locally with individual heterogeneity in preferences or ability, something we will return to below. But the additional structure can make for more natural identifying restrictions than in the case of a program that is at least intended to be uniform, such as a training course. A series of papers has addressed this question in the context of evaluation of Connecticut's welfare-to-work program, Jobs First, against the then-prevailing alternative welfare program. For example, to assess the degree of heterogeneity in treatment responses, [Bitler et al. \(2006\)](#) implement a QTE analysis as described above and relate the resulting estimates to prediction from a standard labor supply model. The [Kline and Tartari \(2016\)](#) study discussed above, aimed at bounding transition probabilities between counterfactual states, takes advantage of across-participant observable differences in the nature of the decision problem faced to construct revealed-preference restrictions on the set of potential transitions. This is an important diagnostic device for assessing the range of counterfactual treatment responses to the program itself. As discussed above, a potential drawback is that the procedure is rather complex and only applies to the particular program studied. One also has to contend with possibly wide bounds.

In principle, Kline and Tartari's approach can also be used for predicting the effect on the distribution of marginal outcomes of moving from traditional welfare to a welfare-to-work program of the same structure at another site (see [Section 3.1](#)). Yet, it is worth keeping in mind that the estimated bounds have the LATE property, i.e., they may depend on the particular distribution of individual characteristics and the local environment. Extrapolating to different populations or environments in their context would require imposing additional assumptions on the underlying static labor supply model and thus trade off additional predictions with robustness.

#### **4.4.2 Addressing the issue *ex ante* through the design of the experiment**

There may be an opportunity to make more progress on this type of treatment effect heterogeneity by building it into the randomization design. Cross-classified and multiple treatment group experiments can be quite helpful for identifying variation in treatment effects.

In some cases, we are directly interested in understanding the distribution of treatment effects. When a plausible structural model (perhaps something as simple as a [Heckman-Vytlacil \(2005\)](#) Roy model) is available, one might use the structural model to predict individual treatment effects and then stratify the experiment based on these predictions. The NIT studies can be seen as a version of this, as these were stratified based on prior earnings, a potentially strong predictor of the treatment effect.

In other cases, concerns about heterogeneity are driven by potential differences between the complier LATE and the population ATE. Rather than simply assigning participants to be offered or not offered the treatment, one might also vary the extent of

efforts to enforce compliance with the experimental assignment. When the relevant selection is thought to be based in part on the anticipated individual treatment effect, as in [Heckman and Vytlacil \(2005\)](#), one can identify the MTE curve directly by randomly assigning participants to multiple values of the incentive (or cost) to obtain the treatment.

Which of these is appropriate depends on the nature of the selection into compliance in the experiment, and how it relates to what would be observed in a nonexperimental setting. To make things concrete, we will consider a study in which applicants are randomly assigned to be eligible or ineligible to receive training offered at a particular job training center. One might expect that noncompliance rates will be low for those assigned to the treatment group for whom it is inconvenient to travel to the program site. One might then expect the LATE to vary with travel costs, but in a simple experiment there is no way to estimate how much of this is due to differences in ATEs between those who live close to and far away from the program site and how much to differences in selection into the complier group.

One way to learn about this would be to implement a more complex, multiple-treatment-arm experiment in which a subset of individuals offered access to the training are also offered transportation to the training site. If the distance—treatment effect curves differ between the two treatment arms, one can conclude that selection into participation is important, and this can then be used (with a parametric selection model) to estimate how the LATE for a similarly selected complier population varies with distance. This may be important if the goal is to generalize from the experiment to a scaled-up program that would offer training at a wider number of sites.

One can also use the three-arm experiment to identify the MTE curve, but only with strong restrictions on the shape of this curve (which correspond to strong parametric assumptions about the selection process; see [Brinch et al., forthcoming](#)). These restrictions may be unattractive. If an important goal of the study is to understand how treatment effects vary with the costs of participation, an even more complex experimental design might be called for. Rather than assigning individuals to a treatment group that receives training at zero cost or a control group that is denied access to training at any price, one might use multiple groups that are offered training at different price points (including potentially negative prices). Variation in outcomes across these groups will trace out several points on the MTE curve and can be used to identify a more flexibly shaped curve under weaker assumptions.

Cross-classified and multiple-treatment-arm experiments raise a number of practical issues that are not confronted in classical treatment/control studies. First, allocating observations across many arms reduces power to detect differences in outcomes between any pair of treatments. Researchers designing experiments must therefore trade off the benefits of a multiple-treatment-arm experiment against reduced ability to detect particular pairwise contrasts. This issue can sometimes be addressed, however, when the alternative arms can be seen as varying the dosage of a single well-defined treatment.

An experiment where all treated individuals are assigned a treatment dose of 1 gives less power for identifying a linear dose–response relationship than one where the same individuals are assigned varying doses with a mean of 1 (for example, when half are assigned a dose of 0.5 and half are assigned 1.5); moreover, the latter design provides at least the chance of detecting nonlinear effects.

Cross-classified experiments, with a fraction  $p$  assigned to treatment A and a fraction  $q$  independently assigned to treatment B, can also be seen as sacrificing power, although again the reality is more complex. Let  $y_{abi}$  represents the potential outcome for individual  $i$  when the program A assignment is  $a$  ( $a = 0$  or 1) and the program B assignment is  $b$ . The traditional estimand for evaluation of program A is  $E[y_{10i} - y_{00i}]$ . Only  $(1 - q)N$  of the  $N$  observations in the cross-classified experiment can be used for estimating this quantity, as the other  $qN$  observations are assigned to receive treatment B. But the experiment has full power for estimating the alternative treatment effect  $E[((1 - q)y_{10i} + qy_{11i}) - ((1 - q)y_{00i} + qy_{01i})]$ . This can be seen as a weighted average of two treatment effects of program A, one that applies to individuals who also receive program B and one for those who do not. In some cases, this may be of more interest than the traditional estimand—e.g., when the scaled-up version of program A will coexist with program B.

## 4.5 Hidden treatments

A long-standing issue in the interpretation of job training program evaluations is that these evaluations commonly have substantial rates of noncompliance and crossovers. Many people assigned to receive training do not complete their courses, and it has been operationally and politically difficult to exclude people assigned to the control group from receiving treatment, either from the same provider that serves the treatment group or from an alternative provider. Indeed, in some cases, ethical concerns led to decisions to actively inform control group individuals about alternative sources of training.

Much of the literature treats this as noncompliance of the type discussed in [Section 2.2.1](#), so estimates the training effect by dividing the ITT effect by an estimate of the complier share (see, e.g., [Heckman et al., 2000](#)). But this is unsatisfactory when the control group noncompliers receive a different treatment—e.g., training from a different provider—from that given to the treatment group. In technical terms, this is a violation of SUTVA; practically, it means that assignment to treatment may affect outcomes even for the always-takers who receive (some type of) training in any case. To our knowledge, this issue has not been addressed in the enormous literature on job training experiments. ([Heckman et al., 2000](#) note the issue, but their analyses focus on nonrandom selection into training and heterogeneity of training effects, which are related but distinct issues.)

Even the instrumental variables (IV) approach, unsatisfactory as it is, is often not feasible: it requires measuring the share of the control group that crosses over. In

many cases, this is not available: the experimental data include information on the receipt of services from the program under study but not on services obtained from other sources. In this case, only ITT estimates can be computed. But these are attenuated by the failure to measure the “hidden” alternative treatments.

#### **4.5.1 Addressing the issue ex post**

A very recent literature takes up this topic in the context of the Head Start preschool program. The Head Start Impact Study randomly assigned Head Start applicants to be offered care or turned away. Many of the control group applicants (and a smaller share of the treatment group) wound up receiving alternative center-based childcare that is thought to be less effective but may be a partial substitute. Where traditional IV estimators treat this as equivalent either to the Head Start treatment or to the receipt of no services, it might be more appropriate to treat it as a distinct, “hidden” treatment.

[Walters \(2014\)](#) estimates heterogeneity in the Head Start effect across centers (sites), finding (among other results) that the LATE of Head Start participation is smaller when more of the complier group is drawn from other centers rather than home-based care. This is suggestive that other center-based care is distinct from home-based care.

[Kline and Walters \(2014\)](#) explicitly model the hidden alternative center treatment, using variation in the compliance patterns across participants’ observable characteristics (e.g., parental education) to identify a multinomial variant of a [Heckman \(1979\)](#) parametric selection correction and thus obtain partially experimental estimates of the separate effects of the two types of childcare. Their approach leverages variation across observable characteristics ( $X$ ) in the share of experimental compliers who are drawn from alternative center care, together with a utility-maximizing choice model that constrains how selection on unobservables varies with  $X$ . With the restrictions imposed by this model, they find large effects of Head Start relative to home-based care. As the Head Start experiment did not directly manipulate the choice between home-based and other center care, they are not able to estimate the relative effect of these with any precision in their least restrictive model, although point estimates are consistent with an effect of other centers comparable to that of Head Start. When Kline and Walters impose stronger restrictions on the selection process, they obtain similar point estimates but with more precision.

[Feller et al. \(2014\)](#) also examine the hidden treatments issue in the Head Start Impact Study sample. They use a principal poststratification approach that, like Kline and Walters, exploits variation across observables in selection into the two treatments. They couple this to a finite mixture modeling strategy that treats the separation of the two complier subgroup distributions as a deconvolution exercise. Parametric assumptions about these distributions are used to identify the LATEs of the two treatments. Results are similar to Kline and Walters: Head Start has positive effects on those who would

otherwise be at home, but little effect on those who would otherwise receive alternative center-based care.

Another example of the analysis of hidden treatments is Pinto's (2015) analysis of the MTO experiment. In one view, the MTO study involved two treatment arms: one offered a housing voucher that could be used anywhere, and the other restricted the voucher to a low-poverty neighborhood. Straightforward experimental comparisons identify the ITT and LATE of usage of each type of voucher. In another view, however, the relevant treatment is the type of neighborhood in which the participant lives. Kling et al., 2007 use variation across the two treatment arms and across sites to identify effects of neighborhood poverty (under restrictions on treatment effect heterogeneity). Pinto (2015) adds more structure, using revealed preference restrictions—anyone offered an unrestricted voucher who moves to a low-poverty neighborhood can be assumed to choose the same type of neighborhood in the counterfactual where he/she receives a restricted voucher—to identify parameters of interest concerning the distribution of neighborhood-type treatment effects.<sup>43</sup>

#### **4.5.2 Addressing the issue *ex ante* through the design of the experiment**

The Pinto (2015) study takes advantage of the multiple-treatment arms in the MTO experiment, while the Head Start papers discussed above exploit, in various ways, the use of centers as strata in that experiment. This suggests, correctly, that complex experimental designs may be useful in resolving hidden treatment problems and that a researcher interested in these problems might be able to design an experiment with them in mind. In the neighborhood effects example, one might want to have several treatment arms that vary in the restrictions they place on neighborhood choice; for Head Start, one might explore a third treatment arm that provides a voucher usable either at a Head Start center or at an alternative center. This design might also be useful for a job training evaluation.

In each of these cases, it is *crucial* to collect information about the type and amount of treatment that each participant actually receives; without this, the complex experimental designs are of little value.

### **4.6 Mechanisms and multiple treatments**

The history in Section 3 makes clear that many labor market experiments involve variation in more than one aspect of a given program. This is clearly the case when programs consisting of suites of services and incentives are evaluated, such as in randomized evaluations of welfare-to-work programs or of large-scale training programs with a range of

<sup>43</sup> Pinto's analysis assumes that the set of neighborhoods in which a voucher can be used is the only relevant difference between the two treatment arms. But in MTO low-poverty voucher recipients were also offered counseling that may have had independent impacts on neighborhood choice or even on outcomes.

integrated services such as JTPA or Job Corps. Yet, even the interpretation of many RCTs of smaller training programs is made difficult by the fact that some form of JSA is provided. Simple RCTs do not identify which of the components of the treatment are responsible for the impact. Learning about such mechanisms, besides being of interest in its own right, is particularly desirable if one wishes to extrapolate to new programs or learn about underlying behavioral parameters. This is, for example, recognized explicitly in the ongoing evaluation of the REA program discussed in [Section 3](#), which aims explicitly at distinguishing the effect of a “hassle” due to being summoned to appear from the actual JSA provided.

Even when the treatment has only one component, in many cases that component is sufficiently complex that the ATE is not enough—we want to understand the underlying mechanism. The simplest example of this is labor supply experiments, for which it is often important to distinguish income and substitution effects. It also arises in many of the welfare reform programs, which can create complex changes in intertemporal budget constraints due to time limits or eligibility effects.

#### **4.6.1 Addressing the issue ex post**

Researchers have used a number of strategies to extract from experimental data evidence on the mechanisms underlying the treatment effects identified by the experiment. In the simplest case, it is sometimes possible to use experimental variation to distinguish the relevant mechanisms, with only minimal restrictions derived from theory. This is most feasible when the experiment involves more than two groups. The first large-scale social experiments, the Negative Income Tax studies, were used in this way. The “treatment” here was a tax schedule described by two parameters: the transfer received if earnings were zero and the tax rate applied to any earnings. The main outcome was labor supply, and a key concern of these studies was to distinguish income from substitution effects.

With a single treatment arm and a single control group, this would not be possible: the net effect of the treatment would be identified, but there would be no way of distinguishing substitution from income effects. (One exception would be if the treatment were designed to be a fully compensated change in the marginal tax rate—this would have no income effect, so the treatment effect would equal the substitution effect. But the NIT treatments were not designed this way.) With multiple treatments that vary both the base transfer and the marginal tax rate, and with an assumption that both income and substitution effects are linear in the relevant tax variable, the two effects can be estimated separately.

To see this, suppose a labor supply function that relates hours of work ( $H$ ) to the wage rate ( $w$ ), nonlabor income ( $N$ ), the marginal tax rate ( $r$ ), and other factors such as preferences for leisure ( $e$ ):

$$H = f(w, N, r, e).$$

For simplicity of exposition, we assume a constant marginal tax rate, although this is not crucial (see [Hausman, 1985](#)). A more restrictive assumption is that the individual labor supply function is linear and additively separable in nonlabor income and the net-of-tax hourly wage:

$$H_i = \gamma_i + w_i(1 - r_i)\delta_i + N_i\eta_i.$$

Now consider a simple experiment that assigns some individuals to a control group where  $r_i$  and  $N_i$  are not manipulated, and others to a treatment group that receives an additional baseline transfer  $D$  and faces an increment to the tax rate  $t$ . Then, adopting the earlier potential outcomes framework, each individual has two potential outcomes:

$$H_{i0} = \gamma_i + w_i(1 - r_i)\delta_i + N_i\eta_i$$

and

$$H_{i1} = \gamma_i + w_i(1 - r_i - t)\delta_i + (N_i + D)\eta_i.$$

With random assignment, the difference in mean labor supply between treatment and control groups equals

$$E[H_i | D_i = 1] - E[H_i | D_i = 0] = -tE[w_i\delta_i] + DE[\eta_i].$$

The first term here represents substitution effects, while the second represents income effects. But the simple experiment identifies only the combination of them.

Fortunately, the NIT studies involved multiple treatment arms, with various combinations of transfers and tax rates. Consider a simple extension of the above structure, with two treatment groups 1 and 2 and associated parameters  $\{D_1, t_1\}$  and  $\{D_2, t_2\}$ . Now each individual has three potential outcomes associated with assignment to the control group and each of the treatment groups,  $H_0$ ,  $H_1$ , and  $H_2$ . Two distinct treatment-control contrasts can be computed:

$$E[H_i | D_i = 1] - E[H_i | D_i = 0] = -t_1E[w_i\delta_i] + D_1E[\eta_i]$$

and

$$E[H_i | D_i = 2] - E[H_i | D_i = 0] = -t_2E[w_i\delta_i] + D_2E[\eta_i].$$

This is a system of two linear equations and two unknowns. So long as the system has full rank—here, as long as  $(D_1/D_2 \neq t_1/t_2)$ —it can be solved for the mean income elasticity of labor supply,  $E[\eta_i]$ , and for  $E[w_i\delta_i]$ . The latter can be divided by the mean wage rate,  $E[w_i]$ , to obtain a wage rate–weighted mean substitution elasticity. (With a large enough sample, the mean substitution elasticity,  $E[\delta_i]$ , could be identified by stratifying the treatment–control comparison by the wage rate.)

A number of studies used the NIT experiment data to estimate the parameters of the labor supply function in basically this way, accounting for additional complications that

we neglect here (e.g., participation decisions, nonlinear tax schedules, etc.) and often using more complex labor supply functions. See, for example, [Moffitt \(1979\)](#). But this was by no means universal: in the late 1970s, the experimental paradigm was not as well developed, and many of the studies that used the experimental data did not rely solely on the randomly assigned components of nonlabor income and tax rates for identification (e.g., [Keeley et al., 1978](#)).

In the above simple model the mean income and labor supply elasticities are just identified with two treatment arms. With more than two arms—the Seattle/Denver experiment alone had 11—the model is overidentified. This opens the possibility of performing overidentification tests of the restrictions imposed when specifying the labor supply function. [Ashenfelter and Plant \(1990\)](#) estimate separate treatment effects of each treatment arm, but we are not aware of studies that investigate formally whether the pattern of effects is consistent with a posited labor supply function.

Even absent multiple treatment arms, sometimes statistical or theoretical models and assumptions can enable researchers to learn about mechanisms that generate a program effect. For example, [Card and Hyslop \(2005\)](#) (henceforth CH) analyze the data from the Canadian SSP RCT. SSP, a welfare-to-work program, combined a strong, temporary work incentive for participating workers with a fixed initial time period during which welfare recipients had to establish eligibility in the program by working full time. As a result of this two-tiered structure, the simple experiment analysis does not distinguish the effects of the various components of the program. This makes it difficult to compare the effects of SSP with other welfare-to-work programs, to assess how SSP worked, and to draw lessons for similar programs. CH use a parametric statistical model to separately identify the effect of the different incentives inherent in the SSP program. In contrast to static evaluations of welfare-to-work programs, CH focus on the dynamic labor supply incentives inherent in the program.

One cannot directly analyze the effect of the subsidy (which in the following we will refer to as the SSP program) for those who became eligible because of selection in the eligibility decision. One can, however, model eligibility as a type of imperfect compliance, permitting the estimation of the LATE of SSP on total employment or on the fraction employed at any given point in time. When one turns to dynamic analyses, potential differential changes in the nature of selection in the treatment and control groups make it impossible to estimate the dynamic responses of hazard rates or wages just based on the RCT.<sup>44</sup> In addition, as in other welfare evaluations, endogenous employment decisions make an analysis of wage outcomes problematic. Another issue is that in the short run the

<sup>44</sup> CH use a standard search theory to model the incentives of SSP and capture the effect of eligibility and the SSP subsidy on labor supply incentives via their effects on the reservation wage. The search model clarifies that, in the presence of heterogeneity, the pool of workers employed at any given point in time may be selected, whether or not there also is sample selection arising from employment decisions (e.g., [Ham and LaLonde, 1996](#)).

strong work incentive arising from the option value in the eligibility period is potentially confounded with the effect of the subsidy.

To address these difficulties, CH proceed by developing a logistic model with random effects and heterogeneity to estimate a benchmark for welfare transitions in the absence of SSP (i.e., for the control group). This model is then combined with parametric specifications of the treatment effects over different ranges of the program spell, as implied by incentives inherent in SSP. This step includes modeling the participation decision and welfare transitions as functions of the SSP subsidy and current and lagged welfare status. A key assumption thereby is that the chosen controls for heterogeneity and the functional form restrictions are sufficient to control for the dynamic selection bias introduced by the eligibility window. CH experiment with different specifications of heterogeneity and provide ample discussion of the goodness of fit of the model. As a result of this exercise, they are able to obtain separate effects of eligibility and SSP. This allows them to simulate the effects of different components of the program and counterfactual policy changes relating to the time path of the subsidy.

The approach and findings in CH suggest that one may not need a structural model to separately identify multiple treatment effects, the dynamic effects of a program, or to simulate the effect of alternative policies. However, an assumption on functional form is required, as well as harder-to-assess assumptions on the form of underlying heterogeneity.

To estimate mechanisms underlying the effect of experimental or policy variation, other papers have used insights from theory to aid identification without estimating a structural model. For example, Schmieder et al. (2016) use insights from the standard search model to estimate the effect of unemployment duration on wages. A recurring question in the analysis and evaluation of welfare and unemployment programs has been the effect of employment and unemployment on productivity and wages. If wages rise with employment duration, welfare-to-work programs can lead to sustained labor force participation. In contrast, if longer nonemployment duration reduces wages, and hence the incentive to work, more generous benefits can lead to a welfare trap.

Card and Hyslop (2005) find that increased employment in the course of the Canadian SSP did little to increase wages. In contrast, Grogger (2005) finds positive wage impacts of employment in the context of a randomized evaluation of Florida's welfare-to-work program.

Few papers have directly analyzed the effect of unemployment duration on wages.<sup>45</sup> The question is difficult for at least two reasons. First, as in Card and Hyslop (2005), even with exogenous variation in incentives at the group level, the type of worker employed at

<sup>45</sup> An exception is Addison and Blackburn (2000), who discuss some of the issues that arise. A larger number of papers have addressed the question of duration dependence in unemployment spells. See Kroft et al. (2013) and references therein.

any given point in the unemployment spell may differ between the treatment and control groups.<sup>46</sup> In other words, it is difficult to find a valid instrument for the duration of unemployment. A second complication arises because even if such variation were available, a change in wages might arise either because of a change in wage offers or due to a change in reservation wages.

To address these difficulties, Schmieder et al. (2016) use the fact that the canonical search model has the strong prediction that forward-looking individuals valuing future UI benefits will respond to a benefit extension by raising their reservation wage well before benefit exhaustion. Unless reservation wages do not bind, this implies that extensions in UI durations should lead to increases in observed reemployment wages throughout the spell. In contrast to this prediction, Schmieder et al. (2016) find in the context of discontinuous increases in UI durations in Germany that reemployment wages at different points of the unemployment spells are unaffected. They deduce that reservation wages likely had little effect on observed wages and hence that the effect of an increase in UI benefit durations on wages arose from an effect of the rise in nonemployment durations on offered wages. In this case, an exogenous increase in UI benefit durations can be used as an instrument to estimate the effect of nonemployment duration on wages.<sup>47</sup>

Another study incorporating theoretical insights from search theory into an empirical study of UI is that of DellaVigna et al. (2016), who analyze a change in the time path of UI benefits in Hungary that kept benefits in the final tier unchanged. They use this variation to structurally estimate key parameters of a model with reference dependence and find the model does quite well compared to an alternative model that explains the pattern based on (unspecified) heterogeneity. The incorporation of nonstandard behavioral assumptions into the evaluation of labor market program is still in its infancy, but is an important avenue for future research.<sup>48</sup>

A closely related topic to the question of mechanisms is the extrapolation of experimental evidence to consider the impacts of new policies, not included in the original evaluation. The value of such extrapolations has long been one of the primary arguments in favor of structural modeling (and against reliance on purely experimental evidence), but some scholars have found out ways to synthesize the approaches. The main challenge here is to bridge between the relatively few parameters that are cleanly identified by an

<sup>46</sup> This bias arises even in the absence of differences in participation.

<sup>47</sup> The authors argue that their test excludes any effect of the worker's outside option on wages, and hence the findings are not specific to the particular model.

<sup>48</sup> For some exceptions, see Lemieux and MacLeod (2000), DellaVigna and Paserman (2005), Oreopoulos (2007); more recently, Chan (2014) examines the role of time inconsistency in the context of the randomized evaluation of Florida Transition Program. Babcock et al. (2012) give an overview of the potential importance of behavioral assumption for the evaluation of public programs.

experiment and the larger set of parameters that are needed to characterize most structural models.

One way to do this is to start with a characterization of structural behavior that is simple enough to be captured within the experimental evidence. For example, if one assumes that the labor supply function is characterized by constant income and (compensated) substitution elasticities, then the estimates of these parameters that are identified by the NIT experiments are sufficient to identify the effects of alternative NIT parameters that were not included in the experimental treatments. A drawback of such an approach is that the range of policies that can be examined is limited. The approach can be extended, of course, to estimate a more complex structural model that either relies on additional statistical and theoretical assumptions, additional nonexperimental moments, or both. In any event, this sort of exercise is on more solid ground when trying to interpolate to values within the range of tax parameters included in the experiment than when these parameters need to be extrapolated outside of that range.

A more recent, closely related approach is known as the “sufficient statistics” approach ([Chetty, 2009](#)). Here, the goal is to characterize optimal policy. Starting with a fully characterized (but usually not overly complex) structural model, it is often possible to derive expressions for social welfare, or for the optimal policy, that depend only on a small number of reduced-form parameters. For example, the Baily-Chetty ([Baily, 1978](#); [Chetty, 2006](#)) formula for optimal UI benefits expresses the optimal benefit level in terms of the elasticity of unemployment duration with respect to UI benefits and the income and substitution effects on the exit hazard from unemployment. If one had experimental evidence regarding these effects, one could use the formula to derive the optimal policy (e.g., [Chetty, 2008](#); [Card et al., 2007](#)).

Of course, any sufficient statistics approach is dependent upon the validity of the underlying structural model—there is no assurance that the true structural model generates the same sufficient statistics as does the one posited by the researcher. In some cases, this may include a relevant class of models and hence provide a degree of robustness. For example, [Chetty \(2009\)](#) gives the example of heterogeneity in treatment effects, where the optimal policy depends only on the mean effect. Yet, it can be hard to know which assumptions in the structural model matter, and generally the assumptions needed to derive the sufficient statistics are fairly strong. At a practical level, conclusions about optimal policies may involve extrapolating very far from the range of policy variation included in the experiment, which means relying strongly on the validity of the theoretical model. In this context, a potential drawback of sufficient statistics is that in contrast to explicitly structural work the empirical fit of the model against the data cannot be assessed.

An alternative approach to obtain a framework for policy extrapolation based on experimental variation is to estimate, or calibrate, a full structural model, using

experimental evidence to aid in identifying (some of) the necessary parameters. One approach is to fix individual parameters at the values indicated by experiments and then calibrate or structurally estimate the remainder. This approach is pursued, for example, by [Davidson and Woodbury \(1997\)](#), who use the Illinois reemployment bonus experiment to estimate the parameters of a search cost function and then combine this function with calibrated values, derived from nonexperimental data, for other parameters of their model of optimal UI benefits. Another approach is to use experimental data to fit a full structural model, but keep the model sufficiently simple such that the main parameters of the model are identified by the available variation, as, for example, in [DellaVigna et al. \(2016\)](#). An alternative is to estimate the structural model solely with nonexperimental data, then use experimental evidence to validate predictions that the model makes for particular reduced-form comparisons (e.g., [Todd and Wolpin, 2006](#)).<sup>49</sup>

#### **4.6.2 Addressing the issue *ex ante* through the design of the experiment**

In some cases, the experimental design can be structured to help uncover the mechanisms underlying the treatment effect of the program. Economic theory may be particularly useful here in connecting fundamental parameters and mechanisms to the types of impacts that can be measured with experiments. One approach is to design an experiment that targets a particular mechanism of interest, rather than identifying the effect of a well-defined program that might be implemented. [Kling et al. \(2017\)](#), this volume, refer to this as a “mechanism experiment,” distinguishing it from a program evaluation. Standard models in labor economics or other fields may provide useful characterizations of the behavioral mechanisms to be tested. For example, models of human capital investment have implications for the factors determining take-up and success of training or schooling programs that may be useful in structuring the experimental design.

A closely related approach is to introduce multiple treatment arms, with program variation among them that can help uncover underlying parameters. The NIT experiments discussed above present a straightforward example of a congenial marriage of classic (static) labor supply theory and the experimental design. As discussed above, as long as both income and substitution effects are linear in the relevant tax measure, multiple treatments manipulating both the base transfer and the marginal tax rate can be used to separately estimate the income and substitution effects.<sup>50</sup>

<sup>49</sup> Another approach to extrapolation that can be viewed as a hybrid between structural and reduced form approaches is to use experimental variation in the incentive to take up a program to effectively estimate a structural model of the compliance rate (e.g., [Heckman and Vytlacil, 2005](#)). As described in [Section 4.4](#), under certain circumstances this allows one to obtain the full distribution of marginal treatment effects and hence to extrapolate.

<sup>50</sup> Multiple treatments may not be necessary. For example, with appropriate data and assumptions, one could in principle experimentally vary *compensated* wage changes to identify the compensated substitution effect. This more closely resembles a mechanism experiment.

The evaluation of the SSP program discussed above is a good example of an experiment that would have benefited from a second treatment arm. Such a treatment might have randomly varied the incentive to become eligible for the (randomly assigned) work subsidy in the main phase of the program. More generally, decisions and programs involving intertemporal trade offs may be an area in which more complex experiments can be particularly insightful. For example, typical UI systems involve expiring benefits, or JSA programs involve sanctions; the timing of benefit exhaustion, reemployment bonuses, or sanctions has been shown to have important empirical effects on reemployment rates (e.g., Meyer, 1995; Black et al., 2003; Schmieder et al., 2012). Hence, experiments that try and get at the underlying behavioral mechanisms may provide important insights into how these programs affect labor supply choices. Knowledge of such mechanisms is also a crucial input in optimizing the delivery of insurance or assistance in the labor market. For example, this could involve a reemployment bonus that declines over time, or one that is available only to those who survive to a specified point. By randomly varying the amount, slope, or intervals, one may gain insights into the nature of intertemporal decision-making relevant for these programs. Intertemporal choice is also an area where theory is likely to be helpful to provide identifying structure. For example, if the goal would be to learn about potential behavioral biases, a model of the effect of particular biases can yield insightful predictions for job search behavior (e.g., DellaVigna et al., 2016).<sup>51</sup>

The usefulness of theory in informing experimental designs hinges, of course, on the model being correct. To mitigate the reliance on particular assumptions (e.g., on functional forms) in principle, one could use revealed preference arguments to generate robust predictions from theory that are then used in design of an experiment. For example, one could use results obtained by Pinto (2015) or Kline and Tartari (2016) to devise multiple treatment arms to test the implied restrictions. However, a model may not be necessary to enrich the experimental design to study underlying channels. The SSP example shows that a basic understanding of the incentives and the nature of the program can be sufficient to design an RCT that uncovers the potentially complex mechanisms underlying the simple SSP evaluation.

## 5. CONCLUSION

Because they allow researchers to control assignment into treatment, RCTs are the Gold Standard for program evaluation. But while random assignment solves the selection

<sup>51</sup> As already mentioned in the discussion of heterogeneous treatment effects, another area where theory is likely to be useful is to understand the determination of compliance rates. As discussed above, the main idea is to experimentally manipulate the incentive to participate and use the variation to trace out the marginal treatment effect (MTE) curve. Theoretical considerations can tell us how to realistically vary the cost of compliance and hence be able to estimate the full range of treatment effects.

problem, there are a broad range of additional relevant design issues that arise routinely in the analysis of central economic questions that are not solved by random assignment on its own. In this chapter, we have discussed six such design issues in depth, including (1) spillover effects and interactions between individuals, leading to a failure of SUTVA; (2) impacts on outcomes that are only observed conditional on individual choices and hence are endogenous, such as wages, hours worked, or participation in a follow-up survey; (3) heterogeneity in treatment effects between experimental sites and observed population groups, or (4) imperfect compliance and heterogeneity in unobserved characteristics, both of which can make it hard to interpret treatment effects and extrapolate to other programs; (5) hidden treatment effects arising because controls also receive versions of the treatment; and (6) the understanding of the mechanisms behind the treatment effect, in particular in the presence of multiple treatment.

We discuss these design issues and solutions in the context of social experiments in the United States labor market, which have provided most of what we know about the functioning of the main labor market programs. Of course, the labor economics literature has been well aware of the limitations of experiments in general and some of these design issues in particular. We have reviewed approaches that can be used to address the design issues in the context of randomized experiments. This includes approaches that can be applied once randomization is completed and ways to modify the experiments itself to address the concerns we identify.

While we discuss design issues in the context of experiments in the labor market, these issues can arise in all areas that have seen active experimental activities, including field experiments discussed elsewhere in this volume. Hence the solutions we identify can be applied to a broad range of questions and should be useful for a wide range of researchers interested in harnessing the power RCTs.

We close with a brief discussion of recent trends in labor market social experiments, several of which highlight the need to pay more attention to the potential design issues in experimental evaluations that we discuss. One overarching trend, cutting across several areas of research, is that academic economists have become more involved with the implementation of experiments. In labor economics, for example, this has meant a shift away from RCTs implemented by large, specialized policy consulting firms (e.g., Mathematica, MDRC, or Abt Associates). For example, several experiments have evaluated take-up of actual government programs within the context of services provided by H&R Block (e.g., [Bettinger et al., 2012](#)). Another example is the increasing number of randomized trials evaluating the role of economic incentives for teachers (e.g., [Fryer et al., 2012; Fryer, 2013; Springer et al., 2010](#)). Similarly, experiments taking place within private businesses have also been quite successful (e.g., [Bandiera et al., 2009](#)).

The greater involvement of academic economists harbors both upside potential, if researchers implement state-of-the-art techniques to address additional design issues, and challenges, as there is a broad range of issues that must be considered and monitored

when implementing an experimental evaluation of an existing program or a new, complex treatment in a real-world setting. We hope the discussion of the design issues in this chapter, as well as our summary of the practical aspects of implementing social experiments, will provide a useful guide for those interested in implementing such social experiments.

A second, related trend has been a movement toward evaluating topics in personnel economics (e.g., the response of teachers to incentive pay programs) as distinct from government social programs. These are often conducted within particular firms and implicate a number of the design issues we discuss, most notably issues of site effects and heterogeneity.

A third important trend has been the use of the actual online labor market, for what amount to field experiments in the taxonomy we set out at the outset (e.g., [Pallais, 2014](#)). The Internet may well provide a useful resource for future social experiments as well. A key advantage may be that researchers may be able to better control the environment, perhaps allowing them to implement more complex study designs that address some of the issues we pose.

## REFERENCES

- Addison, J.T., Blackburn, M.L., 2000. The effects of unemployment insurance on postunemployment earnings. *Labour Econ.* 7 (1), 21–53.
- Ahn, H., Powell, J.L., 1993. Semiparametric estimation of censored selection models with a nonparametric selection mechanism. *J. Econ.* 58 (1), 3–29.
- Alcott, H., 2015. Site selection bias in program evaluation. *Q. J. Econ.* 130 (3), 1117–1165.
- Altonji, J.G., Blank, R.M., 1999. Race and gender in the labor market. *Handb. Labor Econ.* 3 (3), 3143–3259.
- Anderson, M., 2008. Multiple inference and gender differences in the effects of early intervention: a reevaluation of the Abecedarian, Perry Preschool, and Early Training projects. *J. Am. Stat. Assoc.* 103 (484), 1481–1495.
- Angrist, J.D., Hull, P., Pathak, P.A., Walters, C., 2015. Leveraging Lotteries for School Value-Added: Testing and Estimation. National Bureau of Economic Research (Working Paper 21748).
- Angrist, J.D., Imbens, G.W., 1995. Two-stage least squares estimation of average causal effects in models with variable treatment intensity. *J. Am. Stat. Assoc.* 90 (430), 431–442.
- Angrist, J.D., Imbens, G.W., Rubin, D.B., 1996. Identification of causal effects using instrumental variables. *J. Am. Stat. Assoc.* 91 (434), 444–455.
- Angrist, J.D., Krueger, A.B., 1999. Empirical strategies in labor economics. *Handb. Labor Econ.* 3, 1277–1366.
- Ashenfelter, O., Ashmore, D., Deschênes, O., 2005. Do unemployment insurance recipients actively seek work? Evidence from randomized trials in four US states. *J. Econ.* 125 (1–2), 53–75.
- Ashenfelter, O., Plant, M.W., 1990. Nonparametric estimates of the labor-supply effects of negative income tax programs. *J. Labor Econ.* 8 (1), S396–S415.
- Athey, S., Imbens, G., 2017. The econometrics of randomized experiments. In: Duflo, E., Banerjee, A. (Eds.), *Handbook of Field Experiments*, vol. 1, pp. 73–140.
- Babcock, L., Congdon, W.J., Katz, L.F., Mullainathan, S., 2012. Notes on behavioral economics and labor market policy. *IZA J. Labor Policy* 1 (2), 1–14.
- Baily, M.N., 1978. Some aspects of optimal unemployment insurance. *J. Public Econ.* 10 (3), 379–402.

- Baird, S., Bohren, A., McIntosh, C., Ozler, B., 2015. Designing Experiments to Measure Spillover Effects, Second Version (Working Paper 15-021). Penn Institute for Economic Research.
- Bandiera, O., Bankaray, I., Rasul, I., 2009. Social connections and incentives in the workplace: evidence from personnel data. *Econometrica* 77 (4), 1047–1094.
- Barnes, M.S., Benus, J., Cooper, J., Dugan, M.K., Kirsch, M.P., Johnson, T., 2014. U.S. Department of Labor Jobs Corps Process Study. Final Report. U.S. Department of Labor. Available at: [http://wdr.dolela.gov/research/keyword.cfm?fuseaction=dsp\\_resultDetails&pub\\_id=2538&mpy=](http://wdr.dolela.gov/research/keyword.cfm?fuseaction=dsp_resultDetails&pub_id=2538&mpy=).
- Barnow, B.S., 2000. Exploring the relationship between performance management and program impact: a case study of the Job Training Partnership Act. *J. Policy Analysis Manag.* 19 (1), 118–141.
- Becerra, R.M., Lew, V., Mitchell, M.N., Ono, H., 1998. Final Report: California Work Pays Demonstration Project, Report of the First Forty-Two Months. School of Public Policy and Social Research. University of California-Los Angeles, Los Angeles.
- Beecroft, E., Lee, W., Long, D., Holcomb, P.A., Thompson, T.S., Pindus, N., O'Brien, C., Bernstein, J., 2003. The Indiana Welfare Reform Evaluation: Five-Year Impacts, Implementation, Costs and Benefits. Abt Associates, Cambridge, MA.
- Bell, S.H., Bloom, H.S., Cave, G., Doolittle, F., Lin, W., Orr, L.L., 1994. The National JTPA Study: Overview: Impacts, Benefits, and Costs of Title II-A. Abt Associates, Cambridge MA.
- Bell, S.H., Orr, L.L., Burstein, N.R., 1987. Evaluation of the AFDC Homemaker-Home Health Aide Demonstrations: Overview of Evaluation Results. Abt Associates, Cambridge MA.
- Benus, J., Yamagata, E.P., Wang, Y., Blass, E., 2008. Reemployment and Eligibility Assessment (REA) Study: FY 2005 Initiative: Final Report. IMPAQ International, pp. 1–173.
- Bertrand, M., Mullainathan, S., 2004. Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. *Am. Econ. Rev.* 94 (4), 991–1013.
- Bettinger, E., Long, B.T., Oreopoulos, P., Sanbonmatsu, L., 2012. The role of application assistance and information in college decisions: results from the H&R Block FAFSA experiment. *Q. J. Econ.* 127 (3), 1205–1242.
- Bitler, M.P., Gelbach, J.B., Hoynes, H.W., 2006. What mean impacts miss: distributional effects of welfare reform experiments. *Am. Econ. Rev.* 96 (4), 988–1012.
- Black, D.A., Galdo, J., Smith, J.A., 2007. Evaluating the worker profiling and reemployment services system using a regression discontinuity approach. *Am. Econ. Rev.* 97 (2), 104–107.
- Black, D.A., Smith, J.A., Berger, M.C., Noel, B.J., 2003. Is the threat of reemployment services more effective than the services themselves? Evidence from random assignment in the UI system. *Am. Econ. Rev.* 93 (4), 1313–1327.
- Bloom, H.S., Hill, C.J., Riccio, J.A., 2005. Modeling cross-site experimental differences to find out why program effectiveness varies. In: Bloom, H.S. (Ed.), Learning More from Social Experiments: Evolving Analytic Approaches. Russell Sage Foundation, pp. 37–74.
- Bloom, D., Kemple, J.J., Morris, P., Scrivener, S., Verma, N., Hendra, R., December 2000. Final Report on Florida's Initial Time-Limited Welfare Program. Manpower Demonstration Research Corporation, New York.
- Bloom, H.S., Orr, L.L., Bell, S.H., Cave, G., Doolittle, F., Lin, W., Bos, J.M., 1997. The benefits and costs of JTPA Title II-A programs: key findings from the National Job Training Partnership Act Study. *J. Hum. Resour.* 32 (3), 549–576.
- Bloom, D., Scrivener, S., Michalopoulos, C., Morris, P., Hendra, R., Adams-Ciardullo, D., Walter, J., 2002. Jobs First: Final Report on Connecticut's Welfare Reform Initiative. Manpower Demonstration Research Corporation.
- Blundell, R., Bozio, A., Laroque, G., 2011. Labor supply and the extensive margin. *Am. Econ. Rev.* 101 (3), 482–486.
- Blundell, R., Dias, M.C., Meghir, C., Reenen, J.V., 2004. Evaluating the employment impact of a mandatory job search program. *J. Eur. Econ. Assoc.* 2 (4), 569–606.
- Brinch, C., Mogstad, M., Wiswall, M., forthcoming. Beyond LATE with a discrete instrument. *J. Political Econ.*
- Buchinsky, M., 1994. Changes in the US wage structure 1963–1987: application of quantile regression. *Econ. J. Econ. Soc.* 62 (2), 405–458.

- Burghardt, J., Schochet, P.Z., McConnell, S., Johnson, T., Gritz, R.M., Glazerman, S., Homrichausen, J., Jackson, R., 2001. Does Job Corps Work? Summary of the National Job Corps Study. Mathematica Policy Research, Princeton, NJ.
- Card, D., Chetty, R., Weber, A., 2007. Cash-on-hand and competing models of intertemporal behavior: new evidence from the labor market. *Q. J. Econ.* 122 (4), 1511–1560.
- Card, D., Hyslop, D.R., 2005. Estimating the effects of a time-limited earnings subsidy for welfare-leavers. *Econometrica* 73 (6), 1723–1770.
- Card, D., Klueve, J., Weber, A., 2010. Active labor market programs: a meta-analysis. *Econ. J.* 120 (548), F452–F477.
- Cave, G., Bos, H., Doolittle, F., Toussaint, C., 1993. JOBSTART. Final Report on a Program for School Dropouts. Manpower Demonstration Research Corp, New York.
- Cerqua, A., Pellegrini, G., 2014. Do subsidies to private capital boost firms' growth? A multiple regression discontinuity design approach. *J. Public Econ.* 109 (C), 114–126.
- Chan, M.K., 2014. Welfare Dependence and Self-Control: An Empirical Analysis. Working Paper. Economics Discipline Group, UTS Business School, University of Technology, Sydney.
- Chetty, R., 2006. A general formula for the optimal level of social insurance. *J. Public Econ.* 90 (10), 1879–1901.
- Chetty, R., 2008. Moral hazard versus liquidity and optimal unemployment insurance. *J. Political Econ.* 116 (2), 173–234.
- Chetty, R., 2009. Is the taxable income elasticity sufficient to calculate deadweight loss? The implications of evasion and avoidance. *Am. Econ. J. Econ. Policy* 1 (2), 31–52.
- Chetty, R., Friedman, J.N., Rockoff, J.E., 2014. Measuring the impacts of teachers I: evaluating bias in teacher value-added estimates. *Am. Econ. Rev.* 104 (9), 2593–2632.
- Chodorow-Reich, G., Karabounis, L., 2016. The Limited Macroeconomic Effects of Unemployment Benefit Extensions. National Bureau of Economic Research (Working Paper 22163).
- Coglianese, J.J., 2015. Do Unemployment Insurance Extensions Reduce Employment? (Working Paper) Mimeo, Harvard University.
- Corson, W., Decker, P., Dunstan, S.M., Kerachsky, S., 1991. Pennsylvania Reemployment Bonus Demonstration. U.S. Department of Labor, Washington, DC. Final Report (Unemployment Insurance Occasional Paper 92–1).
- Corson, W., Long, D., Nicholson, W., 1984. Evaluation of the Charleston Claimant Placement and Work Test Demonstration. Mathematica Policy Research.
- Crépon, B., Duflo, E., Gurgand, M., Rathelot, R., Zamora, P., 2013. Do labor market policies have displacement effects? Evidence from a clustered randomized experiment. *Q. J. Econ.* 128 (2), 531–580.
- Davidson, C., Woodbury, S.A., 1997. Optimal unemployment insurance. *J. Public Econ.* 64 (3), 359–387.
- Deaton, A., 2010. Instruments, randomization, and learning about development. *J. Econ. Literature* 48 (2), 424–455.
- Dehejia, R.H., Wahba, S., 2002. Propensity score-matching methods for nonexperimental causal studies. *Rev. Econ. Statistics* 84 (1), 151–161.
- DellaVigna, S., Lindner, A., Reizer, B., Schmieder, J.F., 2016. Reference-Dependent Job Search: Evidence from Hungary. National Bureau of Economic Research (Working Paper 22257).
- DellaVigna, S., Paserman, M.D., 2005. Job search and impatience. *J. Labor Econ.* 23 (3), 527–588.
- DiNardo, J., Fortin, N.M., Lemieux, T., 1996. Labor market institutions and the distribution of wages, 1973–1992: a semiparametric approach. *Econometrica* 64 (5), 1001–1044.
- Dorsett, R., Hendra, R., Robins, P.K., Williams, S., 2013. Can Post-Employment Services Combined with Financial Incentives Improve Employment Retention for Welfare Recipients? Evidence from the Texas Employment Retention and Advancement Evaluation. NIESR. Discussion Paper No. 409.
- Farber, H.S., Silverman, D., Wachter, T., 2015. Factors Determining Callbacks to Job Applications by the Unemployed: An Audit Study. National Bureau of Economic Research (Working Paper 21689).
- Fein, D.J., Beecroft, E., Blomquist, J.D., 1994. Ohio Transitions to Independence Demonstration. Final Impacts for JOBS and Work Choice. Abt Associates, Cambridge, MA.
- Feller, A., Grindal, T., Miratrix, L.W., Page, L.C., 2014. Compared to What? Variation in the Impacts of Early Childhood Education by Alternative Care-Type Settings (Working Paper).

- Ferracci, M., Jolivet, G., van den Berg, G.J., 2010. Treatment Evaluation in the Case of Interactions Within Markets. Institute for the Study of Labor (IZA) (No. 4700). Working Paper.
- Fraker, T., Maynard, R., 1987. The adequacy of comparison group designs for evaluations of employment-related programs. *J. Hum. Resour.* 22 (2), 194–227.
- Freedman, S., Friedlander, D., Riccio, J., 1994. GAIN: Benefits, Costs, and Three-Year Impacts of a Welfare-to-Work Program. Manpower Demonstration Research Corp.
- Freedman, S., Knab, J.T., Gennetian, L.A., Navarro, D., 2000. The Los Angeles Jobs-First GAIN Evaluation: Final Report on a Work First Program in a Major Urban Center. Manpower Demonstration Research Corporation, New York.
- Fryer, R., 2013. Teacher incentives and student achievement: evidence from New York City public schools. *J. Labor Econ.* 31 (2), 373–427.
- Fryer, R., Levitt, S.D., List, J., Sadoff, S., 2012. Enhancing the Efficacy of Teacher Incentives Through Loss Aversion: A Field Experiment. National Bureau of Economic Research (Working Paper 18237).
- Gautier, P.A., Muller, P., Rosholm, M., Svarer, M., van der Klaauw, B., 2012. Estimating Equilibrium Effects of Job Search Assistance (No. 9066). CEPR (Discussion Papers).
- Gold, S.F., 1971. The failure of the Work Incentive (WIN) program. *Univ. Pa. Law Rev.* 119 (3), 485–501.
- Greenberg, D.H., Robins, P.K., 1986. The changing role of social experiments in policy analysis. *J. Policy Analysis Manag.* 5 (2), 340–362.
- Greenberg, D.H., Shroder, M., 2004. The Digest of Social Experiments, third ed. The Urban Institute.
- Greenberg, D.H., Shroder, M., Onstott, M., 1999. The social experiment market. *J. Econ. Perspect.* 13 (3), 157–172.
- Grogger, J., 2005. Welfare reform, returns to experience, and wages: using reservation wages to account for sample selection bias. *Rev. Econ. Statistics* 91 (3), 490–502.
- Gronau, R., 1973. The effect of children on the housewife's value of time. *J. Political Econ.* 81 (2), S168–S199.
- Grossman, J.B., Roberts, J., 1989. Welfare savings from employment and training programs for welfare recipients. *Rev. Econ. Statistics* 71 (3), 532–537.
- Gueron, J., 2017. The politics and practice of social experiments: seeds of a revolution. In: Duflo, E., Banerjee, A. (Eds.), *Handbook of Field Experiments*, vol. 1, pp. 27–70.
- Hagedorn, M., Karahan, F., Manovskii, I., Mitman, K., 2015a. Unemployment Benefits and Unemployment in the Great Recession: The Role of Macro Effects. Federal Reserve Bank of New York. Staff Report 646, revised February 2015.
- Hagedorn, M., Manovskii, I., Mitman, K., 2015b. The Impact of Unemployment Benefit Extensions on Employment: The 2014 Employment Miracle? (Working Paper 20884) National Bureau of Economic Research.
- Ham, J.C., LaLonde, R.J., 1996. The effect of sample selection and initial conditions in duration models: evidence from experimental data on training. *Econ. J. Econ. Soc.* 64 (1), 175–205.
- Ham, J.C., Li, X., Reagan, P.B., 2011. Matching and semi-parametric IV estimation, a distance-based measure of migration, and the wages of young men. *J. Econ.* 161 (2), 208–227.
- Hamilton, G., Freedman, S., Gennetian, L., Michalopoulos, C., Walter, J., 2001. National Evaluation of Welfare-to-Work Strategies: How Effective Are Different Welfare-to-Work Approaches? Five-Year Adult and Child Impacts for Eleven Programs. US Department of Health and Human Services and US Department of Education, Washington, DC.
- Hamilton, G., Scrivener, S., 2012. Increasing Employment Stability and Earnings for Low-Wage Workers: Lessons from the Employment Retention and Advancement (ERA) Project. Office of Planning, Research and Evaluation Report 2012-19. Administration for Children and Families. U.S. Department of Health and Human Services.
- Harrison, G.W., List, J.A., 2004. Field experiments. *J. Econ. Literature* 42 (4), 1009–1055.
- Hausman, J.A., 1985. The econometrics of nonlinear budget sets. Fisher-Shultz lecture for the Econometric Society, Dublin: 1982. *Econometrica* 53 (6), 1255–1282.
- Hausman, J.A., Wise, D.A., 1979. Attrition bias in experimental and panel data: the Gary Income Maintenance Experiment. *Econometrica* 47 (2), 455–473.
- Heckman, J.J., 1979. Sample selection bias as a specification error. *Econometrica* 47 (1), 153–161.

- Heckman, J.J., 2010. Building bridges between structural and program evaluation approaches to evaluating policy. *J. Econ. Literature* 48 (2), 356–398.
- Heckman, J., Hohmann, N., Smith, J., Khoo, M., 2000. Substitution and dropout bias in social experiments: a study of an influential social experiment. *Q. J. Econ.* 115 (2), 651–694.
- Heckman, J.J., Hotz, V.J., 1989. Choosing among alternative nonexperimental methods for estimating the impact of social programs: the case of manpower training. *J. Am. Stat. Assoc.* 84 (408), 862–874.
- Heckman, J.J., LaLonde, R.J., Smith, J.A., 1999. The economics and econometrics of active labor market programs. *Handb. Labor Econ.* 3, 1865–2097.
- Heckman, J.J., Smith, J.A., 1995. Assessing the case for social experiments. *J. Econ. Perspect.* 9 (2), 85–110.
- Heckman, J.J., Smith, J., Clements, N., 1997. Making the most out of programme evaluations and social experiments: accounting for heterogeneity in programme impacts. *Rev. Econ. Stud.* 64 (4), 487–535.
- Heckman, J.J., Vytlacil, E., 2005. Structural equations, treatment effects, and econometric policy evaluation. *Econometrica* 73 (3), 669–738.
- Herrem, J.W., Schmitt, L.C., 1983. Eligibility Review Pilot Project Handbook. Wisconsin Department of Industry, Labor, and Human Relations, Madison, WI.
- Holland, P.W., 1986. Statistics and causal inference. *J. Am. Stat. Assoc.* 81 (396), 945–960.
- Horowitz, J.L., Manski, C.F., 2000. Nonparametric analysis of randomized experiments with missing covariate and outcome data. *J. Am. Stat. Assoc.* 95 (449), 77–84.
- Hotz, J., 1992. Recent experience in designing evaluations of social programs: the case of the National JTPA study. In: Garfinkel, I., Manski, C. (Eds.), *Evaluating Welfare and Training Programs*. Harvard University Press, Cambridge, MA, pp. 76–114.
- Hotz, J., Imbens, G., Klerman, J., 2006. Evaluating the differential effects of alternative welfare-to-work training components: a reanalysis of the California GAIN program. *J. Labor Econ.* 24 (3), 521–566.
- Jackson, K.C., Rockoff, J.E., Staiger, D.O., 2014. Teacher effects and teacher-related policies. *Annu. Rev. Econ.* 6 (1), 801–825.
- Jacobson, L.S., 2009. Strengthening One-Stop Career Centers: Helping More Unemployed Workers Find Jobs and Build Skills. Hamilton Project Discussion Paper 2009-01, April. The Brookings Institution, Washington DC.
- Jaggers, M., 1984. ERP Pilot Project Final Report. Wisconsin Department of Industry, Labor, and Human Relations, Madison, WI.
- Johnson, T.R., Pfiester, J.M., West, R.W., Dickinson, K.P., 1984. Design and Implementation of the Claimant Placement and Work Test Demonstration. SRI International, Menlo Park, CA.
- Johnson, W., Kitamura, Y., Neal, D., 2000. Evaluating a simple method for estimating black-white gaps in median wages. *Am. Econ. Rev.* 90 (2), 339–343.
- Johnston, A.C., Mas, A., 2015. Potential Unemployment Insurance Duration and Labor Supply: The Individual and Market-Level Response to a Benefit Cut. Unpublished Working Paper. Princeton University.
- Kane, T.J., Staiger, D.O., 2008. Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation (Working Paper 14607). National Bureau of Economic Research.
- Kane, M.P., 2010. Structural vs. atheoretic approaches to econometrics. *J. Econ.* 156 (1), 3–20.
- Keeley, M.C., Robins, P.K., Spiegelman, R.G., West, R.W., 1978. The estimation of labor supply models using experimental data. *Am. Econ. Rev.* 68 (5), 873–887.
- Kehler, K.C., 1977. The Gary Income Maintenance Experiment: Summary of Initial Findings. Indiana University, p. 91.
- Kemple, J.J., Friedlander, D., Fellerath, V., 1995. Florida's Project Independence. Benefits, Costs, and Two-Year Impacts of Florida's JOBS Program. Manpower Demonstration Research Corporation, New York.
- Kershaw, D., Fair, J., 1976. The New Jersey Income Maintenance Experiment. In: Operations, Surveys and Administration, vol. 1. Academic Press, New York.
- Klepinger, D.H., Johnson, T.R., Joesch, J.M., Benus, J.M., 1997. Evaluation of the Maryland Unemployment Insurance Work Search Demonstration (Unemployment Insurance Occasional Paper 98-2). U.S. Department of Labor, Employment and Training Administration, Unemployment Insurance Service, Washington DC.

- Klepinger, D.H., Johnson, T.R., Joesch, J.M., 2002. Effects of unemployment insurance work-search requirements: the Maryland experiment. *Industrial Labor Relat. Rev.* 56 (1), 3–22.
- Klerman, J.A., Minzner, A., Harkness, J., Mills, S., Cook, R., Savidge-Wilkins, G., 2013. Design report: impact evaluation of Reemployment and Eligibility Assessment Program. Abt Assoc. May 7.
- Kline, P., Tartari, M., 2016. Bounding the labor supply responses to a randomized welfare experiment: a revealed preference approach. *Am. Econ. Rev.* 106 (4), 972–1014.
- Kline, P., Walters, C., 2014. Evaluating Public Programs with Close Substitutes: The Case of Head Start. UC Berkeley Institute for Research on Labor and Employment. Working Paper #123-14.
- Kling, J.R., Liebman, J.B., Katz, L.F., 2007. Experimental analysis of neighborhood effects. *Econometrica* 75 (1), 83–119.
- Kling, J.R., Ludwig, J., Congdon, B., Mullainathan, S., 2017. Social policy: mechanism experiments and policy evaluations. In: Duflo, E., Banerjee, A. (Eds.), *Handbook of Field Experiments*, vol. 2, pp. 389–426.
- Knox, V.W., Miller, C., Gennetian, L.A., 2000. Reforming Welfare and Rewarding Work: A Summary of the Final Report on the Minnesota Family Investment Program, vol. 8. Manpower Demonstration Research Corporation, New York.
- Kornfeld, R., Bloom, H.S., 1999. Measuring program impacts on earnings and employment: do unemployment wage reports from employers agree with surveys of individuals? *J. Labor Econ.* 17 (1), 168–197.
- Kroft, K., Lange, F., Notowidigdo, M.J., 2013. Duration dependence and labor market conditions: evidence from a field experiment. *Q. J. Econ.* 128 (3), 1123–1167.
- Krueger, A.B., Mueller, A.I., 2016. A contribution to the empirics of reservation wages. *Am. Econ. J. Econ. Policy* 8 (1), 142–179.
- LaLonde, R.J., 1986. Evaluating the econometric evaluations of training programs with experimental data. *Am. Econ. Rev.* 76 (4), 604–620.
- Landais, C., Michaillat, P., Saez, E., 2015. A Macroeconomic Theory of Optimal Unemployment Insurance. National Bureau of Economic Research (Working Paper 16526).
- Lee, D.S., 2009. Training, wages, and sample selection: estimating sharp bounds on treatment effects. *Rev. Econ. Stud.* 76 (3), 1071–1102.
- Lemieux, T., MacLeod, W.B., 2000. Supply side hysteresis: the case of the Canadian unemployment insurance system. *J. Public Econ.* 78 (1), 139–170.
- List, J.A., Rasul, I., 2011. Field experiments in labor economics. *Handb. Labor Econ.* 4 (4), 103–228.
- Maguire, S., Freely, J., Clymer, C., Conway, M., Schwartz, D., 2010. Tuning in to Local Labor Markets: Findings from the Sectoral Employment Impact Study. Public/Private Ventures, New York.
- Manpower Demonstration Research Corporation Board of Directors, 1980. Summary and Findings of the National Supported Work Demonstration. Ballinger Publishing Company, Cambridge, MA.
- Meyer, B.D., 1995. Lessons from the US unemployment insurance experiments. *J. Econ. Literature* 33 (1), 91–131.
- Mihaly, K., MaCaffrey, D.F., Staiger, D.O., Lockwood, J.R., 2013. A Composite Estimator of Effective Teaching. Met Project. Available at: [http://www.metproject.org/downloads/MET\\_Composite\\_Estimator\\_of\\_Effective\\_Teaching\\_Research\\_Paper.pdf](http://www.metproject.org/downloads/MET_Composite_Estimator_of_Effective_Teaching_Research_Paper.pdf).
- Miller, C., Van Dok, M., Tessler, B.L., Pennington, A., 2012. Strategies to Help Low-Wage Workers Advance: Implementation and Final Impacts of the Work Advancement and Support Center (WASC) Demonstration. Manpower Demonstration Research Corp, New York.
- Minnesota Department of Jobs and Training, 1990. Re-employ Minnesota (Unemployment Insurance Occasional Paper 90-2). In: Johnson, E.R. (Ed.), *Reemployment Services to Unemployed Workers Having Difficulty Becoming Reemployed*. U.S. Department of Labor, Employment and Training Administration, Unemployment Insurance Service, Washington, DC.
- Moffitt, R.A., 1979. The labor supply response in the Gary experiment. *J. Hum. Resour.* 14 (4), 477–487.
- Newey, W., Powell, J.L., Walker, J.R., 1990. Semiparametric estimation of selection models: some empirical results. *Am. Econ. Rev.* 80 (2), 324–328.

- O'Leary, C.J., 2006. State UI job search rules and reemployment services. *Mon. Labor Rev.* 129 (6), 27–37. <http://research.upjohn.org/jrnarticles/3>.
- Oreopoulos, P., 2007. Do dropouts drop out too soon? Wealth, health and happiness from compulsory schooling. *J. Public Econ.* 91, 2213–2229.
- Pallais, A., 2014. Inefficient hiring in entry-level labor markets. *Am. Econ. Rev.* 104 (11), 3565–3599.
- Palmer, J.L., Pechman, J.A., 1978. Welfare in Rural Areas: The North Carolina-Iowa Income Maintenance Experiment. Brookings Institution, Washington, DC.
- Perez-Johnson, I., Moore, Q., Santillano, R., 2011. Improving the Effectiveness of Individual Training Accounts: Long-Term Findings from an Experimental Evaluation of Three Service Delivery Models. Mathematica, Inc. Final Report.
- Pinto, R., 2015. Selection Bias in a Controlled Experiment: The Case of Moving to Opportunity. Mimeo., University of Chicago.
- Poe-Yamagata, E., Benus, J., Bill, N., Carrington, H., Michaelides, M., Shen, T., 2011. Impact of the Reemployment and Eligibility Assessment (REA) Initiative. Impaq International.
- Powell, J.L., 1984. Least absolute deviations estimation for the censored regression model. *J. Econ.* 25 (3), 303–325.
- Robins, P.K., 1985. A Comparison of the labor supply findings from the four negative income tax experiments. *J. Hum. Resour.* 20 (4), 567–582.
- Rothstein, J., 2010. Teacher quality in educational production: tracking, decay, and student achievement. *Q. J. Econ.* 125 (1), 175–214.
- Rothstein, J., 2016. Revisiting the Impacts of Teachers. Unpublished working paper. [http://eml.berkeley.edu/~jrothst/workingpapers/rothstein\\_cfr.pdf](http://eml.berkeley.edu/~jrothst/workingpapers/rothstein_cfr.pdf).
- Schmieder, J.F., von Wachter, T., Bender, S., 2012. The effects of extended unemployment insurance over the business cycle: evidence from regression discontinuity estimates over 20 years. *Q. J. Econ.* 127 (2), 701–752.
- Schmieder, J.F., von Wachter, T., Bender, S., 2016. The effect of unemployment benefits and nonemployment durations on wages. *Am. Econ. Rev.* 106 (3), 739–777.
- Schochet, P.Z., Burghardt, J.A., 2008. Do Job Corps performance measures track program impacts? *J. Policy Analysis Manag.* 27 (3), 556–576.
- Schochet, P., Burghardt, J., McConnell, S., 2008. Does Job Corps work? Impact Findings from the National Job Corps Study. Mathematica Policy Research.
- Smith, J.A., Todd, P.E., 2005. Does matching overcome LaLonde's critique of nonexperimental estimators? *J. Econ.* 125 (1), 305–353.
- Spiegelman, R.G., O'Leary, C.J., Kline, K.J., 1992. The Washington Reemployment Bonus Experiment. U.S. Department of Labor, Washington, DC. Final report (Unemployment Insurance Occasional Paper 92–6).
- Springer, M.G., Ballou, D., Hamilton, L.S., Le, Vi-N., Lockwood, J.R., McCaffrey, D.F., Pepper, M., Stecher, B.M., 2010. Teacher pay for performance: experimental evidence from the project on incentives in teaching. In: Conference Paper, National Center on Performance Incentives.
- SRI International, 1983. Final Report of the Seattle-Denver Income Experiment. In: Design and Results, vol. I. U.S. Department of Health and Human Services, Washington, DC.
- Steinman, J.P., 1978. The Nevada Claimant Placement Program. Employment Security Research. Nevada Employment Security Department.
- Todd, P.E., Wolpin, K.I., 2006. Assessing the impact of a school subsidy program in Mexico: using a social experiment to validate a dynamic behavioral model of child schooling and fertility. *Am. Econ. Rev.* 96 (5), 1384–1417.
- US Department of Health, Education, and Welfare, 1976. Summary Report: Rural Income Maintenance Experiment. Government Printing Office, Washington, DC.
- Vytlačil, E., 2002. Independence, monotonicity, and latent index models: an equivalence result. *Econometrica* 70 (1), 331–341.
- Walters, C., 2014. Inputs in the Production of Early Childhood Human Capital: Evidence from Head Start (Working Paper 20639). National Bureau of Economic Research.

- Watts, H.W., Rees, A., 1977a. The New Jersey Income Maintenance Experiment. In: *Labor Supply Responses*, vol. II. Academic Press, New York.
- Watts, H.W., Rees, A., 1977b. The New Jersey Income Maintenance Experiment. In: *Expenditures, Health, and Social Behavior, and the Quality of the Evidence*, vol. III. Academic Press, New York.
- Woodbury, S.A., Spiegelman, R.G., 1987. Bonuses to workers and employers to reduce unemployment: randomized trials in Illinois. *Am. Econ. Rev.* 77 (4), 513–530.

# INDEX

‘Note: Page numbers followed by “f” indicate figures and “t” indicate tables.’

## A

- Abdul Latif Jameel Poverty Action Lab (J-PAL), 429  
Abecedarian Project, 112  
Abt Associates, 561  
AC teachers. *See* Alternatively certified teachers (AC teachers)  
Accelerated Math (AM), 131  
Accelerated Reader (AR), 130–131  
Accelerator, 394–395  
ACT. *See* Artemisinin Combination Therapy (ACT)  
Adolescence (AD), 174  
Adult health impacts on productivity, 18–23  
Adulthood (AH), 174  
Advance EITC, 401–402  
AEA. *See* American Economic Association (AEA)  
AEA Randomized Trial Registry, 16  
AES. *See* Agricultural Experiment Station (AES)  
African immigrants in Spain, 96  
Age theory, 104–105  
Agents of government authority, 492  
    incentives  
        for policing and justice, 493  
        for tax collection, 492–493  
Agricultural Experiment Station (AES), 461  
Agricultural Technology Adoption Initiative (ATAI), 429  
Agriculture, 428  
    and FEs, 437–441  
    FEs in, 429–437  
        contracts, 435–436  
        credit markets, 431–432  
        experiments in agriculture, 429–436  
        heterogeneity in conditions, 436  
        index-based weather insurance, 432–434  
        information about agricultural technologies, 429–431  
        price information to farmers, 434–435  
        price response function, 434  
        to reveal production function in agriculture, 460–463  
    spatial dimension of, 444–447

## AH. *See* Adulthood (AH)

- Air pollution, 23  
Alternatively certified teachers (AC teachers), 141–142  
AM. *See* Accelerated Math (AM)  
American Economic Association (AEA), 17  
Andhra Pradesh Randomized Evaluation Studies (APRESt), 361  
ANM. *See* Assistant Nurse Midwives (ANM)  
Antiretroviral therapy (ARV therapy), 20, 55–56, 59–60, 64  
Applicant pool, financial incentives effects on, 485–487  
APRESt. *See* Andhra Pradesh Randomized Evaluation Studies (APRESt)  
AR. *See* Accelerated Reader (AR)  
Artemisinin Combination Therapy (ACT), 44  
ARV therapy. *See* Antiretroviral therapy (ARV therapy)  
ASPE. *See* Assistant Secretary for Policy and Evaluation (ASPE)  
Asset accumulation, market failures preventing, 537–542  
    building long-term financial assets—pensions, 542  
    building productive assets, 538–542  
Assistant Nurse Midwives (ANM), 74  
Assistant Secretary for Policy and Evaluation (ASPE), 561  
ATAI. *See* Agricultural Technology Adoption Initiative (ATAI)  
Average treatment effect (ATE), 565–566, 568, 601, 614

## B

- BAM program. *See* “Becoming a Man” program (BAM program)  
Baseline survey, 525  
BDH program. *See* *Bono de Desarrollo Humano* program (BDH program)  
BDM mechanism, 37–38  
“Becoming a Man” program (BAM program), 138  
    intervention, 419  
    policy evaluations, 420

- Behavioral constraints, 517–518, 530–537  
 adding conditions to incentivize behavior, 534–537  
 providing in-kind transfers to constrain spending choices, 530–534
- Behavioral intervention, 539–540
- Bias in health behaviors, 64–67, 65f
- “Big data” sources, 14–15
- Biometric/time stamp studies, 502
- “Black-box”  
 experiment, 352  
 policy evaluation, 411
- Bono de Desarrollo Humano* program (*BDH* program), 536–537
- Bottom-up approach, 77–79. *See also* Top-down approach
- Broken-windows policing, 391–392, 395
- Budget constraints, 539–540
- C**
- CA. *See* Conservation Agriculture (CA)
- CAB. *See* Circumstances at birth (CAB)
- CAL. *See* Computer aided learning (CAL)
- Calculation and Applied Problems, 137
- California Achievement Test, 5th Edition (CAT), 141–142
- California GAIN program, 582
- California Standardized Testing and Reporting program (STAR program), 120–121
- Canadian Self-Sufficiency Program (SSP), 569
- Candidate mechanism, 400–401, 412–413, 419  
 multiple, 415–416
- CARES. *See* Committed Action to Reduce and End Smoking (CARES)
- Cash transfer experiments, 47–49
- CASL. *See* Classroom Assessment of Student Learning (CASL)
- CAT. *See* California Achievement Test, 5th Edition (CAT)
- CBT. *See* Cognitive behavioral therapy (CBT)
- CCDP. *See* Comprehensive Child Development Program (CCDP)
- CCTs. *See* Conditional cash transfers (CCTs)
- CDDRE. *See* Center for Data-Driven Reform in Education (CDDRE)
- Center for Data-Driven Reform in Education (CDDRE), 158
- Center-based experiments, 110–112
- CH experiment, 624
- Characterizing heterogeneity, 446
- Charter schools, 165–169
- correlation of effect size and average age of intervention, 104f  
 evaluation, 166–169  
 number of students, 165f
- Chicago Public Schools (CPS), 163
- Chicago’s Gautreaux program, 98
- Child health impacts  
 and nutrition on education, 23–28  
 and nutrition on later outcomes, 28–31
- Child Outcomes Study, 126–127
- Chinese Di-Bao Program, 526
- Chlorine, 46, 60–61
- Circumstances at birth (CAB), 174
- Civil-service teachers, 355
- Classroom Assessment of Student Learning (CASL), 153
- Clorin, 50
- Cluster-randomized trial, 33–34
- Clustering, 443
- CM. *See* Control mean (CM)
- CNLSY. *See* NLSY79 Child and Young Adult survey (CNLSY)
- Cognitive behavioral therapy (CBT), 419
- Colombia, 28
- Commitment experiments of health, 67–68
- Committed Action to Reduce and End Smoking (CARES), 67
- Community  
 community-based methods, 522  
 contract, 78
- Compliance  
 imperfect experiments, 566–567  
 perfect experiments with, 564–565
- Compliers, 412
- Comprehensive Child Development Program (CCDP), 115
- Comprehensive Employment and Training Act, 583
- Comprehensive Test of Phonological Processes (CTOPP), 137
- Computer aided learning (CAL), 339, 341–342
- Computer-adaptive learning system, 341–342
- Conditional cash transfers (CCTs), 49, 62, 331, 538–539  
*LCT, vs.*, 537  
 program, 525–526, 530, 534  
 trade-off with UCT, 535–536

- Conservation Agriculture (CA), 429–431
- Contact hypothesis. *See* Intergroup contact theory
- Contexts
- concerns across, 356–357
  - policy-relevant, 398
  - role in moderating policy effects, 405–407
- Contract teachers, 345
- Contracts, 435–436
- Control group, 405–406
- Control mean (CM), 120
- Cost
- data, 328
  - of delays, 449–450
- Cost-sharing terms, 408
- CPS. *See* Chicago Public Schools (CPS)
- Credibility revolution, 560–561
- Credit experiments, 47–49
- Credit markets, 431–432
- Criminology, 421
- Cross-classified experiments, 617–618
- Cross-cutting designs and interactions, 366–367
- CTOPP. *See* Comprehensive Test of Phonological Processes (CTOPP)
- Culture of poverty, 127
- D**
- Data on production and consumption, 451–452
- Data collection, 15, 527–528
- analysis, 372–376
    - cost effectiveness, 376
    - heterogeneity, 374–376
    - mechanisms, 376
    - treatment effects, 372–374
  - intermediate inputs, processes, and mechanisms, 370–371
  - long-term follow-up, 371–372
  - outcomes, 369–370
- Decentralization, 72–73
- Decision-making, 35
- Delivery method for HIV information, 51–52
- Demand, interactions of, 546
- Demand-side interventions, 330–334
- Department of Education (DOE), 145–147
- Department of Labor (DOL), 609–610
- Developing countries, field experiments in
- education in, 330–349, 357–376
  - data collection, 369–372
  - demand-side interventions, 330–334
  - design, 357–367
- evidence, 348–349
- experiment design, 362–367
- governance, 343–348
- implementation and follow up, 368–369
  - intervention design, 358–362
  - pedagogy, 338–343
  - randomization, 368
  - sampling and representativeness, 367–368
  - school and student inputs, 334–338
- Developing economies, 428
- agriculture in, 437
- Deworming, 7–8
- DFS. *See* Double Fortified Salt (DFS)
- Diagnostic test, 55
- Differential effect of public training program, 586
- Disaster relief, 529
- Dispersion, 443
- Displacement, 421
- Distribution scheme, 72
- District-level contracting, 72–74
- DOE. *See* Department of Education (DOE)
- DOJ. *See* US Department of Justice (DOJ)
- DOL. *See* Department of Labor (DOL)
- Double Fortified Salt (DFS), 53
- Double-blinding, 18–19
- double-blind trials useful in agriculture, 459–460
- Droga5 (advertising firm), 133
- E**
- E-governance, 505–507
- Early childhood experiments (EC experiments), 102, 110–115, 174, 183t–201t. *See also* Home environment
- center-based experiments, 110–112
  - home-based experiments, 112–115
  - meta-analysis, 115
- Early Colleges, 134
- Early Start to Emancipation Preparation-Tutoring program (ESTEP-Tutoring program), 139–140
- Early Training Project, 112–113
- Earned Income Tax Credit (EITC), 401, 406
- EAs. *See* Enumeration areas (EAs)
- EC experiments. *See* Early childhood experiments (EC experiments)
- Econometric approach, 10
- Economic(s)
- development, 4, 325–326
  - theory, 403, 627

- Education  
 capital, 325–326  
 Education Production Function, 99  
 IES, 397  
 policy, 121–122  
 value of experiments in education and reasons for growth, 328–329
- Educational attainment, 57–58
- Effectiveness trials, 417–418
- Efficacy trials, 7, 417–418
- Efficiency trials, 7
- EGAP. *See* Experiments in Governance and Politics (EGAP)
- EITC. *See* Earned Income Tax Credit (EITC)
- Eligibility effect, 577–578
- ELL. *See* English Language Learner (ELL)
- Employment Retention and Advancement project (ERA project), 577
- Endogenously observed outcomes, 567–568, 600–607  
 addressing issue ex ante, 606–607  
 addressing issue ex post, 602–606  
 non-parametric selection corrections, 603–606  
 parametric selection corrections, 602–603  
 semi-parametric selection corrections, 603–606
- Enforcement of conditions, 537
- English Language Arts (ELA). *See* Reading
- English Language Learner (ELL), 104
- Enumeration areas (EAs), 49
- Environments  
 determinants of health, 31–34  
 externalities, 453–454
- Epidemiology, 9–10
- Equilibrium effects, 13–14
- ERA project. *See* Employment Retention and Advancement project (ERA project)
- ESSA. *See* Every Student Succeeds Act (ESSA)
- ESTEP-Tutoring program. *See* Early Start to Emancipation Preparation-Tutoring program (ESTEP-Tutoring program)
- Every Student Succeeds Act (ESSA), 121–122
- Ex ante policy evaluation, 397–398
- Ex post policy evaluation, 397–398
- Ex-post policy evaluation, 397–398
- Experimental design, 362–367
- Experimental evaluation, 336
- Experimental evidence, 343, 403
- Experimental methods, 324–325
- Experiments in Governance and Politics (EGAP), 17
- External validity, 353–357, 596  
 concerns across contexts, 356–357  
 implementer heterogeneity, 353–354  
 political economy, 355–356  
 representativeness of study samples, 353  
 varying intervention details, 354–355
- Externalities  
 beware of, 8–14  
 effects, 452–456
- Extreme dosage arm, 395
- F**
- Family Investment Program (FIP), 577
- Family Transition Program, 604
- Field experiments (FEs), 325, 390, 428, 562. *See also* Markets  
 in agriculture, 429–437  
 contracts, 435–436  
 credit markets, 431–432  
 experiments in agriculture, 429–436  
 heterogeneity in conditions, 436  
 index-based weather insurance, 432–434  
 information about agricultural technologies, 429–431  
 price information to farmers, 434–435  
 price response function, 434
- agriculture and, 437–441
- in education  
 alternative framing of questions of interest, 329–330
- in developing countries, 330–349, 357–376
- education and human capital, 325–326
- financing and producing education, 326
- research questions, 326–328
- value of experiments in education and reasons for growth, 328–329
- finding and evaluation method for, 105–110  
 calculation of standard errors, 110  
 Hedge's *g* statistic, 107–110  
 meta-analysis, 108t–109t  
 paper accounting, 107t  
 searches of known databases, 106  
 WWC, 105–106
- implications for design and implementation  
 dependence on random weather realizations and risk, 441–444

- market failures and nonseparability, 450–452  
 measurement, 456–460  
 seasonality and long lags, 447–450  
 spatial dimension of agriculture, 444–447  
 spillovers, externalities, and general equilibrium effects, 452–456  
 limitations, 349–357  
 external validity, 353–357  
 interpreting zero effects, 351–352  
 production function *vs.* policy parameters, 349–351  
 to reveal production function in agriculture, 460–463
- Financial incentives, 71–72, 130–133, 492–497  
 for agents of government authority, 492–493  
 effects on applicant pool, 485–487  
 effects on recruitment, 487–488  
 for frontline service providers, 493–497  
 in primary schools, 130–132  
 in secondary schools, 132–133
- Fine particulate matter (PM2. 5), 23
- FIP. *See Family Investment Program (FIP)*
- Flypaper effects, 47
- Food deserts, 413–414
- Forecast policy effects, 398
- Frontline service providers, incentives for, 493–497  
 health, 494–495  
 incentives on service delivery, 495–496  
 provider attendance, 496–497  
 test scores, 494
- G**  
 “Gain” and “Loss” groups, 148–149  
 Gary Income Maintenance Experiment, 124  
 GDP. *See Gross domestic product (GDP)*  
 GDP per capita  
 public sector  
   age difference, 476–478, 479f  
   education difference, 478–480, 480f–481f  
   gender difference, 476–478, 479f  
   health benefit premium, 476, 478f  
   pay premium, 476, 476f  
   pension premium, 476, 478f  
   tenure premium, 476, 479f  
 GE effects, 454  
 using market surveys to informing, 455  
 measuring cost side of, 455–456
- powering FE to measuring, 455  
 sustaining and scaling-up FE to measuring, 456
- GED. *See General education diploma (GED)*
- General education diploma (GED), 402
- General Education in Manpower Training experiment, 578
- General equilibrium effects, 452–456, 547–548
- General professional development, 151–154
- Generasi program, 76
- Geographic targeting, 521, 523–524
- Geospatial information, 13
- “Gold standard”, 18–19, 80
- GORT. *See Gray Oral Reading Tests-Third Edition (GORT)*
- Governance, 343–348
- Government monitoring  
 collecting information, 503  
 information on performance improve outcomes, 500–502
- Governments screening strategies, 488–491
- Graduation approach, 539
- Graduation programs, 48
- Gray Oral Reading Tests-Third Edition (GORT), 137
- Gross domestic product (GDP), 428
- Group incentives, 145–147
- GSO, 563
- Guided Choice, 586
- H**  
 H&R Block intervention, 413  
 Hb. *See Hemoglobin (Hb)*
- HCD. *See Human capital development (HCD)*
- Head Start, 111–112, 114
- Head Start Impact Study, 619–620
- Health framing, 63
- Health levels in low-income countries  
 demand for health products and healthcare, 34–68  
 bias in health behaviors, 64–67  
 commitment experiments, 67–68  
 incentive experiments, 61–62  
 information experiments, 49–57  
 liquidity experiments, 47–49  
 non-monetary cost experiments, 58–61  
 pricing experiments, 36–47  
 psychology experiments, 62–64  
 schooling experiments, 57–58

- Health levels in low-income countries (*Continued*)
- environmental/infrastructural determinants of health, 31–34
  - experimental estimation of health impact, 18–31
  - methodological section, 6–18
  - experimenting to estimates impacts of health improvements, 8–14
  - experimenting to understanding health behavior determinants, 14–16
  - research transparency, registration, and PAP's, 16–18
  - supply of health care, 69–82
    - experimental audit studies, 69–72
    - improving quality of informal providers, 79–82
    - monitoring experiments, 72–79
  - Hedge's *g* statistic, 107–110
  - Hemoglobin (Hb), 19
  - Herpes simplex virus type 2 (HSV2), 16, 49
  - Heterogeneity, 374–376, 444–447
    - in conditions, 436
    - implementer, 353–354
    - in treatment effect and external validity, 613–618
      - addressing issue ex ante, 616–618
      - addressing issue ex post, 613–616
  - Hidden treatments, 618–620
    - addressing issue ex ante, 620
    - addressing issue ex post, 619–620
  - High-dosage tutoring, 136–139
  - HIV/AIDS, 51
  - Home educational resources, 120–122
  - Home environment, 116–129, 202t–219t. *See also* Early childhood experiments (EC experiments)
  - home educational resources, 120–122
  - meta-analysis, 129
  - neighborhood quality, 127–129
  - parental involvement, 116–119
    - incentives, 118–119
    - information, 117–118
  - poverty reduction experiments, 123–127
    - tax reform, 123–124
    - work programs, 124–127
  - Home-based experiments, 102, 112–115
  - “Hot spots”, 421
  - Household surveys, evidence from, 472–482
    - household survey microdata from countries, 472, 509–511
  - Mincerian returns to education, public sector premium, 481f
  - public sector
    - age difference by GDP per capita, 479f
    - education difference by GDP per capita, 480f–481f
    - gender difference by GDP per capita, 479f
    - health benefit premium by GDP per capita, 478f
    - job benefits on, 477t
    - log pay on, 474t–475t
    - pay premium by GDP per capita, 476f
    - pension premium by GDP per capita, 478f
    - tenure premium by GDP per capita, 479f
  - Household *vs.* farm as unit of analysis, 451
  - Household-bargaining constraints, 517–518
  - Household-specific market failure
    - characterization, 452
  - Housing and Urban Development (HUD), 405–406
  - Houston, evidence from randomized field experiments in, 169–181
    - curriculum, 283t–306t
    - life-cycle model, 182t
      - interpreting literature through, 174
    - simulating impacts on income, 175–181
    - simulating potential impact of implementing best practices in education, 172–181
    - simulating social genome model, 174–175
    - variables across life stages, 176t–180t
  - HSV2. *See* Herpes simplex virus type 2 (HSV2)
  - HUD. *See* Housing and Urban Development (HUD)
  - Human capital, 325–326
    - production, 130, 160
    - spillovers, 326–327
  - Human capital development (HCD), 126–127, 563
  - Hypothetical policy evaluation, 394
- I**
- ICC. *See* Item-Characteristic Curves (ICC)
  - IDA. *See* Iron deficiency anemia (IDA)
  - 3ie. *See* International Initiative for Impact Evaluation (3ie)
  - IES. *See* Institute for Education Sciences (IES)
  - IME. *See* Income Maintenance Experiment (IME)
  - Immunization camps, 59
  - Implementer heterogeneity, 353–354

- In-kind programs  
evaluation, 532–534  
rationale for, 530–532
- In-kind transfers, 530  
to constrain spending choices  
evaluating in-kind programs, 532–534  
rationale for in-kind programs, 530–532
- Incentives, 54  
evidence, 499  
experiments of health, 61–62  
financial, 71–72, 130–133, 492–497  
for agents of government authority, 492–493  
effects on applicant pool, 485–487  
effects on recruitment, 487–488  
for frontline service providers, 493–497  
in primary schools, 130–132  
in secondary schools, 132–133  
financial incentives, 130–133  
in primary schools, 130–132  
in secondary schools, 132–133  
for frontline service providers, 493–497  
health, 494–495  
incentives on service delivery, 495–496  
provider attendance, 496–497  
test scores, 494  
to improving performance, 491–499  
non-financial incentives and returns to schooling, 133–135  
nonfinancial incentives, 497–499  
in parental involvement, 118–119  
for policing and justice, 493  
for tax collection, 492–493  
teacher incentives, 145–151  
efficacy enhancement through framing, 148–149  
group incentives, 145–147  
individual incentives, 147–148  
talent transfers, 150–151
- Incentivize behavior, adding conditions to, 534–537  
enforcing conditions, 536–537  
evaluating conditions relative to basic redistributive programs, 534–536  
imposing conditions, 536
- Income Maintenance Experiment (IME), 556–557, 560, 562, 570–571
- Index-based weather insurance, 432–434
- Indiana Welfare Reform Evaluation, 592
- Individual incentives, 147–148
- Individual productivity, experimental estimation of health impact on, 18–31  
adult health and nutrition on productivity, 18–23  
child health and nutrition on education, 23–28  
child health and nutrition on later outcomes, 28–31
- Individual Training Account experiment (ITA experiment), 586
- Indonesia's Data Collection on Social Protection Programme (PPLS), 523
- Indoor air pollution, 31
- Infectious diseases, 4
- Informal providers (IPs), 80  
improving quality of, 79–82
- Information  
about agricultural technologies, 429–431  
experiments of health, 49–57  
impact on health behavior change, 51–54  
impact on willingness to pay, 49–51  
impact of tailored information on behavior change, 54–57  
impact of targeted information, 57  
flows and monitoring, 500, 504–505  
in parental involvement, 117–118  
on program performance matter, 504–505  
provision, 79  
sharing, 444
- Informing policy, 395
- Infotainment intervention, 53–54
- Infrastructural determinants of health, 31–34
- Input-based incentives, 72–75
- Insecticide-treated bed nets, 13
- Insecticide-treated mosquito nets (ITN), 13
- Institute for Education Sciences (IES), 397
- Insurance market, missing, 517, 528–529
- Intention to treat (ITT), 564  
effect, 564, 566  
estimates, 107–110, 120–121, 149
- Intermittent preventive treatment (IPT), 27
- Internal validity, 596
- International Initiative for Impact Evaluation (3ie), 17
- International standardized tests, 97
- Interventions, 102–103  
BAM, 419  
behavioral, 539–540  
correlation of effect size and average age, 104f  
demand-side, 330–334  
design, 358–362

- I**
- Interventions (*Continued*)
    - H&R Block, 413
    - infotainment, 53–54
    - “participation only”, 79
    - policy, 397–398, 413–414
    - randomized evaluations of, 329–330
    - “relative risk” information, 51
    - student-based
      - financial incentives, 130–133
      - non-financial incentives and returns to schooling, 133–135
      - tutoring, 135–140
    - “summer books”, 418
    - teacher-based, 140–157
      - increasing teacher supply, 141–145
      - teacher incentives, 145–151
      - teacher professional development, 151–157
    - IPs. *See* Informal providers (IPs)
    - IPT. *See* Intermittent preventive treatment (IPT)
    - Iron deficiency anemia (IDA), 18, 20
    - Iron supplementation, 13
    - IRT. *See* Item Response Theory (IRT)
    - Issue ex ante through design of experiment
      - endogenously observed outcomes, 606–607
      - hidden treatments, 620
      - mechanisms and multiple treatments, 627–628
      - site and group effects, 612
      - spillover effects and stable unit treatment value assumption, 599–600
      - treatment effect heterogeneity and external validity, 616–618
    - Issue ex post
      - endogenously observed outcomes, 602–606
        - non-and semiparametric selection corrections, 603–606
        - parametric selection corrections, 602–603
      - hidden treatments, 619–620
      - mechanisms and multiple treatments, 621–627
      - site and group effects, 608–612
      - spillover effects and stable unit treatment value assumption, 597–599
      - treatment effect heterogeneity and external validity, 613–616
    - ITA experiment. *See* Individual Training Account experiment (ITA experiment)
    - Item count technique. *See* List randomization
    - Item Response Theory (IRT), 369
    - Item unmatched count technique. *See* List randomization
    - Item-Characteristic Curves (ICC), 369
    - ITN. *See* Insecticide-treated mosquito nets (ITN)
    - ITT. *See* Intention to treat (ITT)

**J**

- J-PAL. *See* Abdul Latif Jameel Poverty Action Lab (J-PAL)
- Job clubs, 592
- Job Corps, 609
  - evaluation, 585
- Job Corps Study, 610–611
  - services, 584
- Job search assistance (JSA), 563, 567, 569, 571–576, 578–582, 587–594, 604
  - randomized controlled trials of, 588t–591t
- Job Training Partnership Act (JTPA), 402
- Jobs First and Family Transition Program, 576
- Jobs Plus, theory of, 415–416
- JSA. *See* Job search assistance (JSA)
- JTPA. *See* Job Training Partnership Act (JTPA)

**K**

- K–12 schools, randomized field experiments in, 129–169, 220t–282t
  - market-based approaches, 161–169
    - charter schools, 165–169
    - school choice, 163–165
    - vouchers, 161–163
  - school management, 157–160
    - class size, 159–160
    - using data to drive instruction, 158–159
    - extended time, 160
  - student-based interventions
    - financial incentives, 130–133
    - non-financial incentives and returns to schooling, 133–135
    - tutoring, 135–140
- teacher-based interventions, 140–157
  - increasing teacher supply, 141–145
  - teacher incentives, 145–151
  - teacher professional development, 151–157
- Kenya Life Panel Survey (KLPS), 30
- “Kitchen-sink” policy evaluations, 415–416
- KLPS. *See* Kenya Life Panel Survey (KLPS)

**L**

- Labelled cash transfer (LCT), 536–537
- Labor force attachment programs (LFA programs), 126–127

- Labor market social experiments, 570–596  
 job search assistance, 587–594  
 labor supply experiments, 570–578  
 practical aspects of implementing social experiments, 594–596  
 training experiments, 578–587
- Labor market(s)  
 experiments, 567–568  
 social experiments in, 560–563
- Labor supply, 563  
 experiments, 570–578  
 IME, 570–571  
 randomized controlled trials of welfare programs, 572t–575t  
 reemployment subsidy experiments, 577–578  
 welfare reform experiments, 571–577  
 theory, 602
- Lake County Public School District, 121
- Large-scale evaluations of federal education programs and policies, 397
- Large-scale government social experiments, 396
- Large-scale social experiment, 556–557
- LATEs. *See* Local average treatment effects (LATEs)
- LCT. *See* Labelled cash transfer (LCT)
- “Lead farmers”, 429–431
- Learning, heterogeneity and, 447
- Lexile Framework, 121–122
- LFA programs. *See* Labor force attachment programs (LFA programs)
- Life-cycle model, 182t  
 interpreting literature through, 174
- Liquidity experiments of health, 47–49
- List randomization, 15–16
- List response. *See* List randomization
- Livestock, valuing, 458–459
- LLINs. *See* Long-lasting antimarial bed nets (LLINs)
- Local average treatment effects (LATEs), 166–169, 566–567
- Long-lasting antimarial bed nets (LLINs), 40–41
- Long-lasting insecticide treated bed net, 44
- Long-term effects, 546–547
- Long-term financial assets—pensions, building, 542
- Long-term follow-up, 371–372
- “Loss” or “Gain” group. *See* “Gain” and “Loss”; groups
- Low-dosage tutoring, 139–140
- Low-poverty voucher group, 405–406  
 treatment, 408
- M**
- “Managed” professional development, 154–156
- Manpower Demonstration Research Corporation (MDRC), 561, 582
- Marginal treatment effect (MTE), 614
- “Marginality index”, 119
- Market-based approaches, 161–169. *See also* School management; Student-based interventions; Teacher-based interventions
- charter schools, 165–169  
 correlation of effect size and average age of intervention, 104f  
 evaluation, 166–169  
 number of students, 165f  
 school choice, 163–165  
 vouchers, 161–163
- Market(s)  
 failures  
 building long-term financial assets—pensions, 542  
 building productive assets, 538–542  
 and nonseparability, 450–452  
 preventing asset accumulation, 537–542  
 level, 454
- Marketing strategies, 41
- Massive online open course (MOOC), 420
- Match Corps, 138
- Matching estimators, 569
- Maximum Choice, 586
- MC life-stage. *See* Middle childhood life-stage (MC life-stage)
- McCall, William, 98–99
- MDG’s. *See* Millennium Development Goals (MDG’s)
- MDRC. *See* Manpower Demonstration Research Corporation (MDRC)
- Mechanism experiments, 391–392, 394–395, 394f, 627. *See also* Field experiments (FEs)
- complementing sources of evidence, 403–405  
 expanding policies, 407–409  
 history, 396–397  
 “moderators”, 397  
 policy evaluations *vs.*, 409–422  
 BAM and CBT, 419–420  
 charter schools, 421–422

- Mechanism experiments (*Continued*)
  - follow-up policy evaluation, 418–419
  - implementation of policy, 417–418
  - multiple candidate mechanism, 415–416
  - policy experiment checklist, 410t
  - productive strategy, 420
  - resources, 417
  - single mechanism, 412–415
  - sufficiency, 411–416
  - randomized experiments, 390
  - resources on parameters, 398–401
  - role of context in moderating policy effects, 405–407
  - ruling out policies, 398, 401–402
- “Mentoring” effect, 419
- Meta-analysis, 108t–109t
  - early childhood experiments, 115
  - home environment, 129
- “Micro-ordeal” mechanism, 61
- “Microfinance revolution”, 431–432
- Middle childhood life-stage (MC life-stage), 174
- Millennium Development Goals (MDG’s), 325–326
- “Million” experiment, The, 133
- Milwaukee Project, 112
- Mindspark* program, 341–342
- Moderating policy effects, 405–407
- “Moderators”, 397
- Monitoring experiments of health, 72–79
  - bottom-up approach, 77–79
  - district-level contracting, 73–74
  - input-based incentives, 74–75
  - output-based incentives, 75–77
- Monitoring mechanisms
  - government monitoring, 500–503
  - improving program performance, 500
  - information flows and monitoring, 500
  - and public service delivery, 500–505
- Monotonicity, 605
- MOOC. *See* Massive online open course (MOOC)
- Moving to Opportunity experiment (MTO experiment), 127–128, 173, 405–406, 607, 620
  - findings, 128–129
  - voucher treatments, 128
- MTE. *See* Marginal treatment effect (MTE)
- MTO experiment. *See* Moving to Opportunity experiment (MTO experiment)
- Multiple candidate mechanism, 415–416
- Multiple-treatment-arm experiment, 617–618
- N**
- National Alliance for Public Charter Schools, 168
- National Evaluation of Welfare-to-Work Strategies (NEWWS), 587, 592
- National Job Training Partnership Act Study (National JTPA Study), 583
- National Jobs Corps Study, 584, 609
- National JTPA Study. *See* National Job Training Partnership Act Study (National JTPA Study)
- National Longitudinal Survey of Youth 1979 (NLSY79), 105, 174–175
- National Rural Health Mission (NRHM), 75
- National Supported Work Demonstration (NSWD), 578–582
- “Natural experiment”, 391
- NCDs. *See* Non-communicable diseases (NCDs)
- NCLB Act. *See* No Child Left Behind Act (NCLB Act)
- Neighborhood quality, 127–129
- New Hope Project, 577
- New Jersey experiment, 556–557
- New Jersey Income Maintenance Experiment, 396
- New Jersey Negative Income Tax Experiment, 562
- New York City’s Summer Youth Employment Program, 584
- NEWWS. *See* National Evaluation of Welfare-to-Work Strategies (NEWWS)
- NGOs. *See* Nongovernmental organizations (NGOs)
- NLSY79 Child and Young Adult survey (CNLSY), 105, 174–175
- NLSY79. *See* National Longitudinal Survey of Youth 1979 (NLSY79)
- No Child Left Behind Act (NCLB Act), 121–122, 141
- Non-communicable diseases (NCDs), 66–67
- Non-experimental data, 558–559
  - studies, 328, 557, 599
- Non-financial incentives, 497–499
  - intrinsic motivation, 498–499
  - and returns to schooling, 133–135
  - transfers and postings, 497–498

- Non-monetary cost experiments of health, 58–61  
 Non-parametric selection corrections, 603–606  
 Non-random attrition, 607  
 Non-volunteer families, 117  
 Noncontributory pension programs, 542  
 Nongovernmental organizations (NGOs), 7, 24, 73, 353–354  
   NGO-led program, 540  
 Nonprofit organizations, 73  
 Norm-referenced tests, 99  
 NREGAs programs, 543–544  
 NRHM. *See* National Rural Health Mission (NRHM)  
 NSWD. *See* National Supported Work Demonstration (NSWD)  
 Nurse-Family Partnership, 112  
 Nutritional supplementation, 28
- O**  
 “Observational” comparisons, 563–564  
 OEO. *See* Office of Economic Opportunity (OEO)  
 Office for Policy Development and Research (OPDR), 561  
 Office of Economic Opportunity (OEO), 556–557, 562  
 Ohio JOBS program, 582  
 Ohio program, 582  
 OHRP. *See* Office for Human Research Protections (OHRP)  
 One-stop centers, 593  
 OPDR. *See* Office for Policy Development and Research (OPDR)  
 Opportunity Scholarship Program (OSP), 161–162  
 “Ordeal mechanism”, 60–61  
 Organizational constraints, 539–540  
 OSP. *See* Opportunity Scholarship Program (OSP)  
 Output-based incentives, 72–73, 75–77
- P**  
 PACES program, 347–348  
 PALS. *See* Phonological Awareness Literacy Screening (PALS)  
 Paper accounting, 107t  
 PAPs. *See* Preanalysis plans (PAPs)  
 Parametric  
   labor supply models, 562  
   selection corrections, 602–603
- Parental involvement, 116–119  
   incentives, 118–119  
   information, 117–118  
 “Participation only” intervention, 79  
 Pay-for-performance incentives, 76  
 PD. *See* Professional development (PD)  
 Peabody Picture Vocabulary Tests, 99, 125–126  
 Pedagogy, 338–343  
 “Peer farmers”, 429–431  
 Pell grant program, 412  
 Pensions, 470  
 Performance pay system, 361  
 Perry preschool  
   experiment, 98–99  
   program, 110–111  
 Personnel economics of state  
   e-governance and other avenues, 505–507  
   using incentives to improving performance, 491–499  
   monitoring mechanisms and public service delivery  
     government monitoring, 500–503  
     improving program performance, 500  
     information flows and monitoring, 500  
   selection and recruitment of public officials, 482–491  
   stylized facts on architecture of state and role of individuals  
     evidence from household surveys, 472–482  
     key features of state, 470–472  
 Philippines’ Pantawid Pamilyang Pilipino Program (PPP), 534  
 Phonological Awareness Literacy Screening (PALS), 113–114  
 PI. *See* Project Independence (PI)  
 PISA. *See* Programme for International Student Assessment (PISA)  
 “Planning fallacy”, 63  
 Plausible mechanism, 404–405  
 Plot-level panels, 460  
 PM2.5. *See* Fine particulate matter (PM2.5)  
 PMT. *See* Proxy-means testing (PMT)  
 Police-community interactions, 400  
 Policing and justice, incentives for, 493  
 Policing policies, 394  
 Policy evaluation(s), 391, 394–395, 398, 401–402, 404  
   mechanism experiments *vs.*, 409–422  
   BAM and CBT, 419–420

- Policy evaluation(s) (*Continued*)  
 charter schools, 421–422  
 follow-up policy evaluation, 418–419  
 implementation of policy, 417–418  
 multiple candidate mechanism, 415–416  
 policy experiment checklist, 410t  
 productive strategy, 420  
 resources, 417  
 single mechanism, 412–415  
 sufficiency, 411–416
- Policy interventions, 397–398, 413–414
- Policy parameters, 349–351, 409
- Policy-makers, 354, 363, 518
- Policy-oriented researchers, 395
- Policy-relevant  
 contexts, 398  
 information, 392–393, 403, 415
- Policy-research funders, 395
- Policymakers, 394
- Political economy, 355–356
- Potential outcomes, 564
- Poverty reduction experiments, 123–127  
 tax reform, 123–124  
 work programs, 124–127
- Power and baselines, 364–365
- PPLS. *See* Indonesia's Data Collection on Social Protection Programme (PPLS)
- PPP. *See* Purchasing power parity (PPP)
- PPPP. *See* Philippines'; Pantawid Pamilyang Pilipino Program (PPPP)
- Pratham's instructional approach, 339–340
- Preanalysis plans (PAPs), 16–18
- Price(s)  
 elasticity, 36  
 information to farmers, 434–435  
 measuring, 457–458  
 response function, 434
- Pricing  
 experiments of health, 36–47  
 family labor, 458  
 methods to estimating demand curve, 36–40  
 TIOLI experiments *vs.* stated WTP, 39f  
 results, 40–47  
 purchase rate of preventive health products, 45f
- Privacy for survey respondents, 15–16
- Private-sector auditors, 503
- Production function, 349–351, 444–445  
 in agriculture, 460–463
- Productive assets building, 538–542
- Professional development (PD), 152–153  
 CASL, 153
- Program evaluation, social experiments as tool for, 563–567  
 experiments with perfect compliance, 564–565  
 imperfect compliance, 566–567  
 local average treatment effect, 566–567
- Programme for International Student Assessment (PISA), 97
- PROGRESA  
 CCT program, 354–355  
 experiment, 118–119
- Progress in International Reading Literacy Study, 97
- Project Independence (PI), 592
- Project STAR, 97–99, 159
- Proxy-means testing (PMT), 521–522
- Psychological factors, 35
- Psychology experiments of health, 62–64
- Psychosocial stimulation, 28
- Public goods, 468–469
- Public health  
 measurement, 31  
 studies, 14–15, 23–24, 31, 69
- Public officials  
 selection and recruitment of, 482–491  
 demographic differences of public sector, 483t–484t  
 financial incentives, 485–488  
 governments screening, 488–491
- Public policies, 412
- Public sector  
 age difference by GDP per capita, 479f  
 demographic differences of, 483t–484t  
 education difference by GDP per capita, 480f–481f  
 gender difference by GDP per capita, 479f  
 health benefit premium by GDP per capita, 478f  
 job benefits on, 477t  
 log pay on, 474t–475t  
 pay premium by GDP per capita, 476f  
 pension premium by GDP per capita, 478f  
 tenure premium by GDP per capita, 479f
- Public service delivery  
 evidence, 505  
 government monitoring, 500–503  
 improving program performance, 500  
 information flows and monitoring, 500

- information flows and monitoring by citizens, 504–505
- Public training program, differential effect of, 586
- Public-sector wage gaps, 508
- Pupil-teacher ratio, 336
- Purchasing power parity (PPP), 540
- Q**
- QTEs. *See* Quantile treatment effects (QTEs)
- Quantile treatment effects (QTEs), 375, 603
- Quasi-experimental evidence, 404  
and structural research designs, 569
- R**
- Racial and ethnic inequality, 96
- RAND health Insurance Experiment, 395, 407–408
- Random assignment, 399–400, 558–559  
experiments, 557  
of students, 97–98
- Randomization, 368, 527  
level and form of, 526–527  
unit, 362–364
- Randomized control trials (RCTs), 100, 324–325, 331, 353–354, 520, 533, 557, 577  
implicit production function, 462  
of JSA, 588t–591t  
of Reading Recovery, 156  
standard RCT empirical approaches, 9–10  
welfare programs, 572t–575t
- Randomized experiments, 390. *See also* Field experiments (FEs)
- Randomized field experiments, 97–98, 390–391  
early childhood experiments, 110–115, 183t–201t  
evidence in Houston, 169–181  
curriculum, 283t–306t  
interpreting literature through simple life-cycle model, 174  
life-cycle model, 182t  
simulating impacts on income, 175–181  
simulating potential impact of implementing best practices in education, 172–181  
simulating social genome model, 174–175  
variables across life stages, 176t–180t
- home environment, 116–129, 202t–219t  
in K-12 schools, 129–169, 220t–282t
- market-based approaches, 161–169
- school management, 157–160
- student-based interventions, 130–140
- teacher-based interventions, 140–157
- Rapid malaria diagnostic tests (RDTs), 55
- Rationale for in-kind programs, 530–532
- RCM. *See* Rubin Causal Model (RCM)
- RCTs. *See* Randomized control trials (RCTs)
- RDP. *See* Regional Development Program (RDP)
- RDTs. *See* Rapid malaria diagnostic tests (RDTs)
- REA system. *See* Reemployment and Eligibility Assessment system (REA system)
- Reading, 122, 135, 172
- Reading Recovery (RR), 155–156
- Real policies, replicating tests of, 398
- REAP. *See* Rural Education Action Program (REAP)
- Recruitment, financial incentives effects on, 487–488
- Redistributive programs, 519–528, 534–536  
evaluating impacts of redistributive programs, 526–528  
targeting poor, 520–526  
experimentally testing between targeting methods, 523–526  
targeting methods, 521–523
- Reemployment and Eligibility Assessment system (REA system), 397, 592–594
- Reemployment subsidy experiments, 577–578
- Regional Development Program (RDP), 485
- Registration, 16–18
- Registry for International Development Impact Evaluations, (RIDIE), 17
- “Relative risk” information intervention, 51
- Research transparency, 16–18
- RIDIE. *See* Registry for International Development Impact Evaluations, (RIDIE)
- Riverside County, 582
- Roma in Hungary, 96
- RR. *See* Reading Recovery (RR)
- Ruling out policies, 398, 401–402
- Rural Education Action Program (REAP), 363
- S**
- Safety net programs, 516–517
- School choice, 163–165, 347–348
- School effectiveness, 168–169

- School input, field experiments in education in developing countries, 334–338
- School management, 157–160. *See also* Market-based approaches; Student-based interventions; Teacher-based interventions
- class size, 159–160
- using data to drive instruction, 158–159
- extended time, 160
- School-based experiments, 102
- Schooling experiments of health, 57–58
- SDAs. *See* Service Delivery Areas (SDAs)
- SDGs. *See* Sustainable Development Goals (SDGs)
- Seasonality and long lags, 447–450
- Seattle/Denver experiment, 623
- Selection-on-observables estimators, 569
- Self Sufficiency Program (SSP), 623, 628
- Self-provided inputs, quantifying, 457
- Self-Sufficiency Program (SSP), 124–125, 543–544, 577
- Self-targeting programs, 522
- Semi-parametric
- labor supply models, 562
  - selection corrections, 603–606
- Service delivery, incentives on, 495–496
- Service Delivery Areas (SDAs), 583
- SGM. *See* Social Genome Model (SGM)
- Single mechanism, 412–415
- Site and group effects, 568, 607–612
- addressing issue ex ante, 612
  - addressing issue ex post, 608–612
- Social desirability bias, 15
- Social experiment. *See also* Field experiments (FEs)
- design issues, 596–628
  - endogenously observed outcomes, 600–607
  - hidden treatments, 618–620
  - mechanisms and multiple treatments, 620–628
  - site and group effects, 607–612
  - spillover effects, 597–600
  - SUTVA, 597–600
  - treatment effect heterogeneity and external validity, 613–618
  - in labor market, 560–563
  - limitations of experimental paradigm, 567–569
  - endogenously observed outcomes, 567–568
  - hidden treatments, 568–569
  - mechanisms and multiple treatments, 569
  - site and group effects, 568
  - spillover effects, 567
  - SUTVA, 567
- treatment effect heterogeneity and external validity, 568
- quasiexperimental and structural research designs, 569
- as tool for program evaluation, 563–567
- experiments with perfect compliance, 564–565
- imperfect compliance, 566–567
- local average treatment effect, 566–567
- Social Genome Model (SGM), 105, 173
- simulation, 174–175
- Social interactions, role of, 414–415
- Social policy
- experiment, 405, 407
  - social policymaking, 399
- Social protection, 528–529
- behavioral constraints, 530–537
  - general equilibrium effects, 547–548
  - implementation matters, 542–545
  - interactions of demand and supply, 546
  - long-term effects, 546–547
  - market failures preventing asset accumulation, 537–542
  - missing insurance markets, 528–529
  - redistributive programs, 519–528
  - evaluating impacts of redistributive programs, 526–528
  - targeting poor, 520–526
- “Softer” commitment devices, 68
- Soil and seed quality, measuring, 459
- SP method. *See* Standardized patient method (SP method)
- Spatial conditions, heterogeneity of, 445–446
- Spatial dispersion of agriculture, 439, 444–447
- Spillover effects, 13–14, 452–456, 567, 597–600
- addressing issue ex ante, 599–600
  - addressing issue ex post, 597–599
- SSP. *See* Canadian Self-Sufficiency Program (SSP); Self Sufficiency Program (SSP); Self-Sufficiency Program (SSP)
- Stable unit treatment value assumption (SUTVA), 565, 567, 597–600
- addressing issue ex ante, 599–600
  - addressing issue ex post, 597–599
- Standardized patient method (SP method), 69–70
- STAR program. *See* California Standardized Testing and Reporting program (STAR program)
- State capacity, 500

- STI tests, 16, 62
- Stratification, design randomized experiments
- Stratified randomized experiments
- Structural approaches, 569
- Structural methods, 559
- Structured Choice, 586
- Student achievement, 116–119, 131, 140, 142, 145–146, 148, 151, 154–155
- Student incentives, 362
- Student input, field experiments in education in developing countries, 334–338
- Student-based interventions. *See also* Market-based approaches; School management; Teacher-based interventions
- financial incentives, 130–133
- in primary schools, 130–132
- in secondary schools, 132–133
- non-financial incentives and returns to schooling, 133–135
- tutoring, 135–140
- high-dosage tutoring, 136–139
- low-dosage tutoring, 139–140
- Subsidies, 9–10, 33–34, 36, 42, 50–51
- “Success for All” program, 154–155
- “Sufficient statistics” approach, 626
- “Sugar daddies”, 51
- “Summer books” intervention, 418
- “Sunk cost” effect, 36
- Supply, interactions of, 546
- Sustainable Development Goals (SDGs), 325–326
- SUTVA. *See* Stable unit treatment value assumption (SUTVA)
- T**
- Take-it-or-leave-it (TIOLI), 37
- TIOLI experiments *vs.* stated WTP, 39f
- Talent Transfer Initiative (TTI), 150–151
- Talent transfers, 150–151
- TAP. *See* Teacher Advancement Program (TAP)
- Targeting methods, 521–526
- TaRL. *See* Teaching at Right Level (TaRL)
- Tax(es), 500
- collection, incentives for, 492–493
- reform, 123–124
- TC teachers. *See* Traditionally certified teachers (TC teachers)
- Teach for America (TFA), 141, 143
- corps members, 144
- teachers, 143–144
- Teacher Advancement Program (TAP), 147–148
- Teacher feedback, 156–157
- Teacher incentives, 145–151, 362
- efficacy enhancement through framing, 148–149
- group incentives, 145–147
- individual incentives, 147–148
- talent transfers, 150–151
- Teacher professional development
- general professional development, 151–154
- “managed” professional development, 154–156
- teacher feedback, 156–157
- Teacher quality, 337
- Teacher-based interventions, 140–157. *See also* Market-based approaches; School management; Student-based interventions
- increasing teacher supply, 141–145
- teacher incentives, 145–151
- teacher professional development, 151–157
- Teaching at Right Level (TaRL), 339
- “Technology of instruction”, 338
- Texas evaluation, The, 577
- TFA. *See* Teach for America (TFA)
- The New Teacher Project (TNTP), 144–145
- Three-arm experiment, 617
- “Timely audits”, 501
- TIMSS. *See* Trends in International Mathematics and Science Study (TIMSS)
- TIOLI. *See* Take-it-or-leave-it (TIOLI)
- TNTP. *See* The New Teacher Project (TNTP)
- Top-down approach
- input-based incentives, 74–75
- output-based incentives, 75–77
- TOT. *See* Treatment on treated (TOT)
- Trade Adjustment Assistance Community College and Career Training Grants Program (TAACCCT Grants Program), 583–584
- Trade Adjustment Assistance program, 409
- Traditional voucher
- group, 405–406
- treatment, 408
- Traditionally certified teachers (TC teachers), 141–142
- Training experiments, 578–587
- Job Corps, 587
- evaluation, 585
- JTPA program, 583
- NSWD study, 582
- randomized controlled trials of programs, 579t–581t

Transaction costs, 444–447  
 Transition to adulthood (TTA), 174  
 “Treatment dose”, 408–409, 414  
 Treatment on treated (TOT), 564–566, 601  
 Trends in International Mathematics and Science Study (TIMSS), 97  
 TTA. *See* Transition to adulthood (TTA)  
 TTI. *See* Talent Transfer Initiative (TTI)  
 Tulsa’s universal pre-kindergarten program, 112  
 Turkish immigrants in Germany, 96  
 Tutoring, 135–140  
     high-dosage, 136–139  
     low-dosage, 139–140  
 Two-way ANOVA, 122

## U

UCTs. *See* Unconditional cash transfers (UCTs)  
 UFT. *See* United Federation of Teachers (UFT)  
 UI. *See* Unemployment insurance (UI)  
 Ultimate policy concern, 394–395  
 Unconditional cash transfers (UCTs), 49, 331,  
     526, 538–539  
     trade-off with CCT, 535–536  
 Unemployment insurance (UI), 528–529,  
     558–559, 577, 587, 592–593, 598, 625  
 United Federation of Teachers (UFT), 145–147  
 United Nations Development Programme, 106  
 Unmatched count. *See* List randomization  
 US Department of Education, 390  
 US Department of Health and Human Services,  
     571–576  
 US Department of Justice (DOJ), 391

## V

Value added tax (VAT), 503  
 Value of experiments in education and reasons for  
     growth, 328–329  
 Value-added model (VAM), 327, 373  
 “Value-added” scores, 612  
 VAM. *See* Value-added model (VAM)  
 VAT. *See* Value added tax (VAT)  
 VCT. *See* Voluntary counseling and testing (VCT)  
 Victim reporting, 400–401  
 Voluntary counseling and testing (VCT), 56–57  
 Volunteer families, 117–118  
 Voucher(s), 161–163  
     approach, 61

## W

Wage inequality, imulating potential impact of  
     implementing best practices in education  
     on, 172–181  
 Weberian model, 470  
 Welfare reforms, 571–577  
 Welfare-to-work experiments, 576–577  
 What Works Clearinghouse (WWC), 100, 100f,  
     105–106  
 WIA. *See* Workforce Investment Act (WIA)  
 Wide Range Achievement Test 3–Spelling  
     (WRAT), 137  
 Willingness to pay (WTP), 36  
 WISE. *See* Work and Iron Status Evaluation  
     (WISE)  
 WJ-R. *See* Woodcock-Johnson Psycho-  
     Educational Battery-Revised (WJ-R)  
 Woodcock Reading Mastery Tests-Revised  
     (WRMT), 137  
 Woodcock-Johnson Psycho-Educational Battery  
     achievement test, 112  
 Woodcock-Johnson Psycho-Educational Battery-  
     Revised (WJ-R), 137  
 Woodcock-Johnson Tests of Achievement, 99  
 Work and Iron Status Evaluation (WISE), 18–19  
 Work programs, 124–127  
 Worker Profiling and Reemployment Services  
     program (WPRS program), 593  
 “Workfare” programs, 517–518  
 Workforce Innovation and Opportunity Act  
     (2013), 584  
 Workforce Investment Act (WIA), 583–584, 586  
 Worm infections, 25  
 WPRS program. *See* Worker Profiling and  
     Reemployment Services program  
     (WPRS program)  
 WRAT. *See* Wide Range Achievement Test  
     3–Spelling (WRAT)  
 WRMT. *See* Woodcock Reading Mastery Tests-  
     Revised (WRMT)  
 WTP. *See* Willingness to pay (WTP)  
 WWC. *See* What Works Clearinghouse (WWC)

## Y

Ypsilanti Perry Preschool Project, 103

## Z

Zero effects interpretation, 351–352