NBER WORKING PAPER SERIES

THE PRODUCTION OF HUMAN CAPITAL IN DEVELOPED COUNTRIES: EVIDENCE FROM 196 RANDOMIZED FIELD EXPERIMENTS

Roland G. Fryer, Jr

Working Paper 22130
http://www.nber.org/papers/w22130

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
March 2016

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, Meghan Howard-Noveck, C. Adam Pfander, and Rucha Vankudre 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 rfryer@fas.harvard.edu. The usual caveat applies. The views expressed herein are those of the author and do not necessarily reflect the views of the National Bureau of Economic Research.

The Production of Human Capital in Developed Countries: Evidence from 196 Randomized Field Experiments

Roland G. Fryer, Jr
NBER Working Paper No. 22130
March 2016
JEL No. I0,J0,J38

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.

Roland G. Fryer, Jr
Department of Economics
Harvard University
Littauer Center 208
Cambridge, MA 02138
and NBER
rfryer@fas.harvard.edu

A Online appendix is available at http://www.nber.org/data-appendix/w22130

"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 twenty-four percent 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 twenty percent less (Hatton 2011). The Roma in Hungary are over two years less educated, have worse self-reported health, and earn twenty-eight percent less (Kántor 2011). Turkish immigrants in Germany are almost twice as likely to be unemployed and earn thirty-eight percent less (von Loeffelholz 2011). African immigrants in Spain are less educated than natives, have a 4.9 percentage point higher unemployment rate, and earn thirty-five percent 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 U.S., O'Neill (1990) and Neal and Johnson (1996) demonstrate that blacks, Hispanics, and whites are paid similar prices for similar pre-market 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 seventy percent.

An important question then, is what obstacles preclude the acquisition of productive skills. Using ten large datasets which together, include students that range in age from eight 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 U.S. fall behind by age two (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 (PIRLS) 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 policy makers 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 non-experimental 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 non-experimental 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 to 82 percent of the white-black test score gap in the data

Similarly, a large body of non-experimental 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, Brooks-Gunn, and Klebanov 1994; Halpern-Felsher et al. 1997; Kohen et al. 2002). However, randomized and quasi-experimental studies have failed to establish a causal link. Although Rosenbuam (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, Oreopoulous (2003) found no evidence of long-term impacts of neighborhood quality on labor market outcomes in a quasi-experimental analysis. More importantly, in the short-run, the Moving to Opportunity randomized housing mobility experiment (Ludwig et al. 2012; Kling, Liebman and Katz 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-twenties, 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.[1] 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 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 alter 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.[2] 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 (EPF) 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

[1] See Levitt and List (2009) for a brief history of field experiments.

[2] 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).

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 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, Levitt, and List 2015). 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, and programs designed to increase the human capital of the adults in school buildings (e.g. Fryer 2014; 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 percent 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 percent. Figure 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's 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 what – together– our collective effort over the past several decades has taught us about how to produce human capital in developed countries.[3]

[3] 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 (2016), the latter in this volume. We describe a few here. First, many questions that are potentially interesting to economists may not be answerable

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 non-laboratory 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 one thousand 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 with standardized reading or math outcomes.[4] 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.[5] Early childhood experiments investigate the impacts of preschool attendance, home-based initiatives that target pre-kindergarten children, and different preschool models on early achievement. Indeed, any experiment with outcomes measured before kids enter school is categorized as early childhood – inde-

---

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 randomized controlled trials (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 U.S. 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 pre-existing 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 the budget of the researcher.

[4] 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.

[5] We don't 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.

pendent 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 meta-analysis coefficient (see DerSimonian and Laird 1986) – the labels on categories become a "lazy man's" way of deciding what works and what doesn't. If the meta-analysis 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).

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 meta-coefficients for early childhood experiments are 0.111σ (0.031) for standardized math scores and 0.165σ (0.032) for reading. The estimates for schools are 0.052σ (0.008) and 0.068σ (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 0.507-0.655σ, 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 0.779-1.582σ, 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 laborius 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 provide 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 don't align with traditional economic choice variables in a concise way and because of the potential effects of publication bias on these types of studies.

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 forty 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 2014; 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 non-standard 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 suggests 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. The 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 0.177σ. The average effect of reading interventions targeting students with an average age greater than 14 is 0.039σ.

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-a-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 U.S. 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 (U.S. Department of Education 2015).[6] 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 2011; 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 most all cases, the randomized studies were already in the What Works Clearhinghouse, 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 phases 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.[7] 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

[6] 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.

[7] 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.

of studies from 859 to 196. As discussed, we only included experiments that had samples determined by a verifiably random process that were pre-college; 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 quasi-experimental 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.

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 non-experimental. Housing demolitions in Jacob (2004) and the famous brown and 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; Cohen et al. 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 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).[8] 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 micro-data, 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 Hedge's 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 five decades there have been many field experiments designed to increase achievement before kids enter school.[9] Appendix Table 1 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 in order to support both the cognitive and non-cognitive development of the children over the course of two years beginning when the children were three years old. Schweinhart, Barnes, and Weikart (1993) find that students in the Perry Preschool program had higher test scores between the ages of 5 and 27, 21 percent less grade retention or special services required, 21 percent higher graduation rates, and half the number of lifetime arrests in comparison to chil-

[8] We documented all assumptions that were made for each study and these can be obtained from the author upon request.

[9] See Carneiro and Heckman (2003) or Almond and Currie (2010) for extensive reviews.

dren 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 percent, 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) investigates 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 find statistically and economically significant impacts.

Another important center-based intervention, which was initiated three 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.[10] 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 $7,300 per child.

Evaluations of Head Start have often been difficult to perform due to the typical non-random nature of enrollment in the program.[11] Puma et al. (2010), in response to the 1998 reauthorization of Head Start, conduct an evaluation using randomized admission into Head Start.[12] The impact of being offered admission into Head Start for 3- and 4-year-olds is 0.10 to 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 to 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

[10] 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 percent of the federal poverty level.

[11] 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 ten. 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.

[12] Students not chosen by lottery to participate in Head Start were not precluded from attending other high-quality early childhood centers. Roughly ninety percent of the treatment sample and forty-three percent of the control sample attended center-based care.

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).[13]

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 pre-kindergarten program. The Abecedarian Project provided full-time, high-quality center-based childcare services for four cohorts of children from low-income families from infancy through age five 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 is two years old – a total of fifty 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 two. 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 fifteen 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 fourth grade. The Infant Health 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 three, as well as enhanced early childhood educational care and bimonthly parent group meetings. Brooks-Gunn,

[13] 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).

Liaw, and Klebanov (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 seven hours a day, five days a week until the children were six years old. Garber (1988) finds that this program resulted in an increase of 23 points on the Stanford-Binet IQ test at age six 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 to 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.[14] They found that children in treatment families score 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, upper-case letter knowledge, beginning word sounds, print and word awareness, rhyme awareness, and 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 upper-case letter subtest and the summed score. There was limited evidence that READY4K! had differential impacts across family characteristics.

[14] ~~Note that since this study did not report results from a norm-referenced outcome, it was not included in our tables and analysis.~~

Fryer, Levitt, and List (2015) 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 which were designed as information sessions to aid parents in educating their children, proof of homework completion, and the performance of their children 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 0.119σ (with a standard error of 0.094). These estimates are non-trivial, but smaller in magnitude than some classroom based interventions. For instance, the impact of Head Start on test scores is approximately 0.145σ. The impact of the Perry Preschool intervention on achievement at 14 years old is 0.203σ. 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, Levitt, and List (2015) report that the impact of being offered a chance to participate in our parental incentive scheme on non-cognitive skills is large and statistically significant (0.203σ (0.083)). These results are consistent with Kautz et al. (2014), who argue that parental investment is an important contributor to non-cognitive 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 pre-determined subsamples that we blocked on experimentally. Two stark patterns appear in the data. The first pattern is along racial lines: Hispanics (48 percent of the sample) and whites (8 percent of the sample) demonstrate large and significant increases in both cognitive and non-cognitive domains. For instance, the impact of the parent academy for Hispanic children is 0.367σ (0.133) on our cognitive score and 0.428σ (0.122) on our non-cognitive score. Among the small sample of whites, the impacts are 0.932σ (0.353) on cognitive and 0.821σ (0.181) on non-cognitive. The identical estimates for blacks are actually negative but statistically insignificant on both cognitive and non-cognitive dimensions: −0.234σ (0.134) and −0.059σ (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, pre-treatment 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 pre-treatment test scores. Students who enter our program below the median on non-cognitive skills see no benefits from our intervention in either the cognitive or non-cognitive domain. In stark contrast, students who enter our parent academy above the median in non-cognitive skills experience treatment effects of roughly 0.3 standard deviations on both cognitive and non-cognitive dimensions. If we segment children by both cognitive and non-cognitive pre-treatment scores, the greatest gains are made on both the cognitive and non-cognitive dimension by students who start the program above the median on non-cognitive 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 impacts 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 0.111σ (0.031) for math interventions and 0.189σ (0.027) for reading. Of the 64 treatment effects recorded in these randomized studies, 21 were statistically positive; zero were statistically negative and 43 were statistically indistinguishable from zero.[15]

[15] 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; Wilson and Matheny 1983; 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 underachievement 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 though welfare-to-work programs and tax reform to moving families to better neighborhoods. The literature is summarized in Appendix Table 2.

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 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.

A. 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 afterschool 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 6th 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 non-volunteer 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, non-volunteers in treatment classes, volunteers in control classes, and non-volunteers in control classes.

The study was interested in addressing three outcomes: (1) parental involvement attitudes and behav-

ior; (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 non-volunteer 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 -0.001σ (0.021) on math scores and 0.034σ (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.

B. 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. Socio-economic data was 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 two 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 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 school. 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 percent. For secondary school students, enrollment increased by 7.2 to 9.3 percentage points for boys and 3.5 to 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, Sengupta, and Todd (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 for 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 two years of treatment. Their analysis is based on 4,800 families with 11,000 children out of which half were assigned to treatment and the other half to control. Opportunity NYC spent $8,700 per family in treatment over three 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 wifi, 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 quasi-experimental data have generated mixed results. Some studies find large positive effects (Attewell and Battle 1999; Fiorini 2010; Schmitt and Wadsworth 2006; Fairlie 2005; Fairlie, Beltran and Das 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 1,123 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 percentage points more likely to have a computer at follow-up, 25 percentage points more likely to have Internet service, they reported using a computer 2.5 hours more

per week than control students' average of 4.2 hours, 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 hours more per week for schoolwork (control mean (CM) was 1.89 hours), 0.42 hours more for e-mail (CM = 0.25 hours), 0.80 hours more for games (CM = 0.84 hours), and 0.57 hours more for social networking (CM = 0.57 hours).

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 percent 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.[16] 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 multi-ethnic 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 multi-racial 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

[16] See Heyns (1978), Cooper et al. (1996), Alexander et al. (2001), Broh (2004), Heyns (1987), Klibanoff & 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".

Left Behind Act.[17] 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 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 if 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 0.08σ (0.04) impact on students' standardized reading test scores and there were differential effects by race. Treatment increased test scores by 0.22σ (0.09) for black students, 0.14σ (0.08) for Latino students, and 0.17σ (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

[17] Note that the No Child Left Behind Act was superseded by the Every Student Succeeds Act in December 2015.

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 0.046σ (0.033) annual impact over the three years of the experiment.

The meta-coefficients for home educational resource experiment were -0.060σ (0.050) for math scores and 0.015σ (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 percent of black children and 37 percent of Hispanic children experience poverty while only 10 percent 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.

A. 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 U.S. 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 1,799 eligible families, 57 percent

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 percent) and two guaranteed income levels (about three fourths of the poverty level or equal to the poverty level). The lower guarantee level was about $1,000 a year more than the support level of the Indiana Aid to Families with Dependent Children program.[18] The tax rate is the amount by which the negative income tax payment is reduced for each dollar of income that a family earns.

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 $5,200 and the national average at that time was $9,433) and over 40 percent 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 experiments (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 $2,000 per year (approximately a 50 percent increase). For married mothers, there was no change in hours worked per week. For female family heads, there was a decrease of about two hours 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.

B. 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 it's western

[18] In 1972, the official poverty level for a four-person non-farm family was $4,275.

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.[19]

The study randomly assigned 6,000 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 hours 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 three years of the program and received more than $18,000 in supplements over that 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 $3,400 more than control members. Total income (supplements, earnings, and income assistance) increased by $6,300 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 six years after random assignment (two 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

[19] Statistics come from the 2011 Canadian census (Statistics Canada 2013).

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 U.S., 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 pre-employment 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 five-year follow-up period, a majority of control group members worked at some point. For example, 88 percent of the control parents from the Grand Rapids site were employed at some point. In Oklahoma City, 79 percent worked at some point during that time and 66 percent 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 five 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 to 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 percent 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, a 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 meta-coefficients for this strand of literature are 0.008σ (0.029) and 0.016σ (0.024). And, perhaps more telling, there is not one experiment that generates statistically significant positive effects on standardized test scores.[20]

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; Harrignton 1982). Taking the culture of poverty paradigm at face value, the randomized field experiment that one would ideally conduct would be to move

[20] However, note that some of these studies found impacts for sub samples of the participants or on non-cognitive outcomes.

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 4,604 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 percent poverty rate in 1990) and some housing-mobility counseling; (2) the Section-8 only voucher group, which received 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 percent of household heads were non-Hispanic blacks, 31 percent were Hispanic, and nearly all households were female-headed at baseline. About half of the experimental voucher group and 63 percent 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 to 7 years after random assignment) and long-term (10 to 15 years after random assignment) follow-up surveys and analyses (Kling, Liebman, and Katz 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 percent poverty rate. MTO led to a 9 percentage point decline in the duration-weighted average tract poverty rate over the 10-15 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 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, Liebman and Katz 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-twenties, 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 old 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 are 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 -0.004σ (0.008) for math interventions and 0.010σ (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 effect the production of human capital of school children. In this section, we explore 105 randomized field experiments conducted in K-12 schools. The literature is summarized in Appendix Table 3.

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

3.3.1 Student-Based Interventions

A. Financial Incentives

Perhaps the most natural way to increase human capital production – at least to an economist – is to change the incentives of school children 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 post-secondary education around the globe (Angrist et al. 2002; Angrist and Lavy 2009; Kremer, Miguel, and Thornton 2009; Behrman, Sengupta, and Todd 2005; Angrist, Bettinger, and Kremer 2006; Angrist, Lang, and Oreopoulos 2009; Fryer 2011; Fryer and Holden 2013; Barrera-Osorio et al. 2011; Bettinger 2012; Hahn, Leavitt, and Aaron 1994; Jackson 2010). We describe a subset of the literature below.

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.[21] Bettinger (2012), who evaluates a pay-for-performance program for students in grades three through six in Coshocton, Ohio, is one notable exception. Coshocton is ninety-four percent white and fifty-five percent free/reduced-price lunch. Students in grades three through six took achievement tests in five different subjects: math, reading, writing, science, and social studies. Bettinger (2012) reports a 0.13σ 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 second 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 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 1,777 children across the 21 treatment schools.

Paying students to read books yielded a treatment effect of 0.012σ (0.069) in reading and 0.079σ (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 0.173σ (0.069) increase in reading achievement among English speaking students and a 0.118σ (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 0.221σ (0.068) and -0.164σ (0.095) for students who are English Language Learners.

Fryer and Holden (2013) conducted a randomized field experiment in fifty traditionally low-performing public schools in Houston, Texas – providing financial incentives to fifth grade students, their parents, and

[21] The use of non-financial 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.

their teachers in twenty-five 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 1,821 parents, and $393,038 to 1,734 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 σ (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 0.081σ (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 -0.077σ (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 1.66σ more objectives, have parents who attend two more parent-teacher conferences, have 0.228σ higher standardized math test scores and equal reading scores relative to high-achieving students in control schools. Conversely, lower-achieving students master 0.686σ more objectives, have parents who attend 1.5 more parent-teacher conferences, have equal math test scores and 0.165σ *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 [0.271σ (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 [-0.219σ (0.084)]. These data suggests that there may be long-run impacts of multitasking through learning, dynamic complementarities, or both.

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 ten interim assessments administered by the NYC Department of Education to all students. In Chicago, ninth graders were paid every five 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) and Fryer (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 0.004σ (0.017) in reading and -0.031σ (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 0.152σ (0.092) in reading and 0.114σ (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., 0.10σ) 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 0.093σ (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 D.C. 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 asks their teacher 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), Fryer (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 second grade students to read books significantly increases reading achieve-

ment 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.

B. Non-Financial Incentives and Returns to Schooling

Fryer (2013) describes a large and innovative randomized field experiment which grew out of a partnership between three large organizations: Tracphone – the largest pre-paid mobile phone provider in the US, 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 pre-loaded 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 non-financial 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 (2013) 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 percent 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 two years of college credit or a two-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 ten 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 eight of the Early Colleges. The final sample included 2,458 students for the administrative outcomes and 1,294 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 percent of lottery losers and 80 percent of Early College students ever enrolled in college whereas only 71 percent 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 six years of data available, the gap four years after 9th grade was 39.2 percentage points and this gap had decreased to 9.8 percentage points six 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 four-year college and they find no significant differences for any college or two-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 percent). These degrees were typically associate's degrees and approximately 20 percent 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 four 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 meta-analysis coefficients for student incentives experiments are 0.024σ (0.018) for math achievement and 0.021σ (0.017) for reading.

C. 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 minutes per week, others are provided 5 hours of intense instruction in the same time period. This heterogeneity leads, naturally, to large differences in treatment effects. Dobbie and Fryer (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 Dobbie and Fryer (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 three days per week or being tutored at a rate that would equate to 50 hours or more over a 36-week period.[22]

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 eleven 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. In order 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 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 minutes a day, five days a week, from September to June. This resulted in the average treatment student attending 126 sessions or 105 hours

[22] ~~We add to the Dobbie and Fryer (2013) definition because not all studies report days and group size.~~

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 hours of training on early childhood interventions, early reading acquisition, and teaching strategies. Additionally, tutors received 2 hours 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 – nine 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 hours 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 non-normed Comprehensive Test of Phonological Processes (CTOPP) four times during the treatment year and four 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 0.55σ on the comprehension subtest of the GORT to 1.69σ 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 0.30σ and 0.24σ on the GORT accuracy and GORT comprehension subtests, respectively). The authors saw similar reading results for the non-standardized 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 -0.33σ and -0.37σ 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 were both black and eligible for free or reduced-price lunch. The intervention consisted of providing students with non-academic 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 one hour 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 in order to participate. For the academic portion of the intervention, students met in groups of two with a math tutor one hour a day, every day. The tutors were hired following the methodology of Match Corps and were paid $16,000 plus benefits for the nine-month academic year.[23] 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). In order 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 0.51σ and increased math GPA by 0.425 grade points on a four-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 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.

[23] 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.

Summarizing, the meta-coefficient on high-dosage tutoring is 0.309σ (0.106) for math achievement and 0.229σ (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.

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 three 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 twelve 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 hours 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 three 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 four months passed between assignment and a student's first meeting with a tutor. Throughout the two years of the study researchers found that 61.8 percent of treatment students eventually participated in ESTEP-Tutoring and the average treatment student received 18 hours of math tutoring and 17 hours 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 non-participants were no longer in the foster home listed on their application. Once the tutoring program was initiated, students were eligible for 50 hours of tutoring delivered through 2 hour sessions twice a week.

In order to investigate the impact of ESTEP-Tutoring on these students, Courtney et al. (2008) conducted three interviews over the two years after randomization (baseline, one year out, and two 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 (two years after random assignment) due to the fact that participation in ESTEP was still ongoing for many students one 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.

Putting all low-dosage tutoring experiments together, the meta-coefficient on low-dosage tutoring is 0.015σ (0.013) for math achievement and 0.015σ (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 0.15σ to 0.24σ in math and 0.15σ to 0.20σ 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 quasi-experimental 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. 2008). 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.

A. 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 (NYCTF) 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.

In order 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 7 states across the nation. Within these schools, 2,610 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 relative 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 different 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 course work 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 non-experimental 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 percent of the variation in effects on math test scores and 1 percent 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 don't 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 non-profit organization that recruits recent college graduates to teach for two years in low-income communities, is one of the nation's 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 2,500 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 8,200 TFA corps members in 125 "high-need" districts across the country, including 13 of the 20 districts with the lowest graduation rates. Roughly 80 percent of the students reached by TFA qualify for free or reduced-price lunch and more than 90 percent are black or Hispanic.

Entry into TFA is highly competitive; in 2010, more than 46,000 individuals applied for just over 4,000 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 complete an online application, which includes a letter of intent and a resume. After a phone 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 five-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 percent 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 $1,500 per corps member fee to TFA to offset screening and recruiting costs. TFA gives corps members various additional financial benefits, including "education awards" of $4,725 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 non-TFA teacher. At the end of school year, Glazerman et al. (2006) found that students assigned to TFA teachers score about 0.12σ higher in math and 0.03σ 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 that did not enter teacher 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 0.07σ 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 two years of teacher 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.

B. Teacher Incentives

To increase teacher productivity, there is growing enthusiasm among policy makers for initiatives that tie teacher incentives to the achievement of their students. Since 2006, the U.S. 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 2013).

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).

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.[24] The experiment was a randomized school-based trial. Each participating school could earn $3,000 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 $1,500 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 sixty 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% of, 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 $3,000 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 -0.018σ (0.024) in mathematics and -0.014σ (0.020) in reading for elementary schools, and -0.046σ (0.018) in math and -0.030σ (0.011) in reading for middle

[24] 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.

schools, *per year*. Thus, if an elementary school student attended schools that implemented the teacher incentive program for three years, her test scores would decline by -0.054σ in math and by -0.042σ in reading, neither of which is statistically significant. For middle school students, however, the negative impacts are more sizeable: -0.138σ in math and -0.090σ in reading over a three-year period.

Consistent with Fryer (2013), 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.

Individual Incentives

Springer et al. (2010) evaluated Tennessee's POINT program – a three-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 $5,000 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 2013; 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 non-trivial 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 (2013) 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 $2,000 per teacher and teachers could earn an additional $7,000 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 sixteen 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 one 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 percent while teachers in control schools had a retention rate of 82.8 percent, 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 -0.04σ (0.05) and the effect size for math was -0.04σ (0.06). The test sore impacts were insignificant across all grade levels and were robust to various sensitivity analyses.

Enhancing the Efficacy of Teacher Incentives Through Framing

During the 2010-2011 and the 2011-2012 school years, Fryer et al. (2015) 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.[25] Other teachers – the "Loss" treatment – were 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 (2011), 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 are minimized.

The first year ITT results of our experiment are consistent with over three 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 (0.455σ (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 0.245σ (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 (2011) 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 0.087σ (0.088) and for "Gain" are 0.115σ (0.109). The pooled estimates for both years of the experiment are 0.210σ (0.069) and 0.116σ

[25] Note that although math, reading, and science test scores were incentivized (the latter only for fourth and seventh 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.

(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 0.796σ (0.209) for "Loss" and 0.376σ (0.168) for "Gain". In the second year, the estimates were 0.574σ (0.176) and 0.714σ (0.144) for "Loss" and "Gain" respectively. Therefore, the pooled effect size for both years and both treatments was 0.568σ (0.121) for kindergarten math scores.

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 percent within their subject and grade) $20,000, paid over two 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. In order to receive the full financial incentive,

high-performing teachers must remain in the low-achieving school for a full two 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. In order to achieve this high rate of transfer, over 1,500 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 four years more experience and were 11 percentage points more likely to have a National Board Certification than control teachers (CM = 9 percent). 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 five 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 percent of the control teachers) and were less likely to receive mentoring (39 compared to 59 percent 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 0.18σ and students' reading scores by 0.10σ in the first year after randomization. In the second year, the cumulative impacts for treatment students were 0.22σ and 0.25σ 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 percent compared to 70 percent). In addition, retention rates after the completion of the second year of the experiment were not statistically significant. Approximately 60 percent of treatment teachers returned to the low-achieving school for a third, non-incentivized school year. In comparison, a statistically indistinguishable 51 percent of control teachers remained in the fall of the third school year.

The meta-coefficient on teacher incentives is 0.022σ (0.022) for math achievement and -0.006σ (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 (2011) and Neal (2011) seriously when designing teacher incentive schemes.

C. Teacher Professional Development

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 (U.S. 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 2010-2012 (U.S. 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).[26] 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 U.S. 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 second 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

[26] A national study revealed that over 80% of elementary and secondary teachers reported participating in 24 hours or less of professional development over the 2005-2006 school year and summer of 2006 (U.S. Department of Education 2009).

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 two years after randomization as a follow-up.

On average, teachers in the first intervention reported attending 39 hours of PD and teachers in the second intervention reported attending 47 hours of PD. In comparison, control teachers only reported attending 13 hours of PD. Garet et al. (2008) found that this exposure to PD lead to significant impacts on teachers' knowledge and practices for both groups. Both interventions had positive impacts on second 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 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.

In order 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 fourth grade and one fifth 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 fifth 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 percent of teachers returned to the research team. In order to combat the alternative hypothesis that any impact was just the result of the intervention schools having more resources, control schools were given $1,000.

In order 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. In order 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 (0.42σ) 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 meta-coefficient for general PD is 0.019σ (0.024) for math achievement and 0.022σ (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).

'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 don't 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 in order to improve overall student achievement and is currently used in 1,200 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 eleven 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 three years after randomization for the baseline cohort of kindergarten students. Although forty-one schools were initially randomized, only thirty-five schools were included in the analysis due to six schools dropping out over the years for various reasons. The authors conclude that these thirty-five 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 0.36σ (0.11) on phonemic awareness, 0.24σ (0.11) on word identification, and 0.21σ (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 every day for a 30-minute lesson over 12 to 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 post-graduate program where they are expected to emerge as literary experts. Literature on RR reports that approximately 75 percent 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, thirty-seven 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-minute 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 0.9σ to 2.02σ and treatment had an impact of 0.94σ on scores from the Slosson Oral Reading Test—Revised.

In 2010, the U.S. 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 2,000 schools and provide literary assistance to over 88,000 students. May et al. (2013) report the findings from the first two years of this scale-up.

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 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) investigate the causal impact of being assigned to the RR program. They find that RR increased student achievement by 0.60σ on the reading words subtest and 0.61σ on the reading comprehension subtest.

The effects of these 'managed' PD experiments for both subjects are statistically significant and for reading, quite large. The meta-coefficient is 0.052σ (0.016) for math achievement and 0.403σ (0.120) for reading achievement.

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. Dobbie and Fryer (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 quasi-experimental 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 eleven percent 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 1,800 high schools from eight countries and student achievement in those schools. They find a strong correlation between higher management quality and better educational outcomes. Dobbie and Fryer (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. Dobbie and Fryer (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 Dobbie and Fryer (2013).

A. 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.

In order 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 0.059σ (0.029) and increased students' reading scores by 0.033σ (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 four through eight were eligible to participate (over 1,000 schools); 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 post-intervention 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.

B. 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 five to seven 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 0.19σ-0.28σ and represent 64 to 82 percent 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.

C. 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 five-week summer program that lasted for 3.5 hours 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. 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 0.69σ 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 twenty elementary schools to either a full or half-day schedule. Holmes and McConnell found that full-day kindergartners had math scores that were 0.29σ lower and reading scores that were 0.11σ 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 5th, 6th, and 7th grade students to participate in an extended school day program. The program consisted of a classroom of approximately ten students receiving an additional two hours of language instruction, two hours of math instruction, and one hour of excursions per week. Meyer and Van Klaveren found that assignment to treatment increased math scores by 0.087σ (0.067) and increased reading scores by 0.005σ (0.081). Neither of these effects are 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.

A. 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 (0.08σ-0.12σ) more per year in math compared to lottery losers. This suggests effect sizes on the order of 0.32σ-0.48σ for four 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 to 2.31 percentile points.

The DC Opportunity Scholarship Program (OSP) is another voucher program that provides up to $7,500 to low-income families in the District of Columbia to send their children to participating private schools. Wolf et al. (2010) use 2,300 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 $1,400 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 three 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 non-participants 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 less 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. Additinoally, 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 three 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 meta-coefficient on voucher experiments is 0.024σ (0.021) for math achievement and 0.030σ (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.[27]

[27] A notable counter example is a recent experiment implemented in India. Muralidharan and Sundararaman (2015) 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 (2015) 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 non-voucher private students. This suggests that vouchers are a cost-effective way to potentially increase student achievement without any negative externalities.

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.

B. School Choice

Cullen, Jacob, and Levitt (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.[28] 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.

The analysis in Cullen, Jacob, and Levitt (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, Jacob, and Levitt (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 non-academic 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 non-traditional outcomes, possibly supporting the hypothesis that students and parents choose choice schools for non-academic reasons. Cullen, Jacob, and Levitt (2006) found that lottery winners report fewer incidents of

[28] 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.

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 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 sixth 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 eighth grade outcomes of lottery winners to lottery losers, Bifulco et al. (2009) find that students offered admission to the magnet schools scored 0.109σ higher on math and 0.252σ higher on reading tests, of which only the latter was statistically significant at conventional levels.

C. 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. Figure 2 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 educational practices that could be used to inform traditional public schools with new ideas and fresh approaches.

In America, charter schools currently enroll almost four percent 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.

Evaluating Charter Schools

The method for evaluating charter schools is remarkably consistent across the literature.[29] 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:

$$outcome_i = \mu + \gamma X_i + \pi Z_i + \sum \nu_j Lottery_{ij} + \sum \phi_j Lottery_{ij} * 1(sibling_i) + \eta_i \quad (1)$$

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 (LEP) services, and a quadratic in two prior years of math and ELA test scores. $Lottery_{ij}$ is an indicator for entering the lottery in year j, and $1(sibling_i)$ indicates whether student i had a sibling enter the lottery in the same year.[30] Equation (1) identifies the impact of *being offered a chance* to attend a charter school, π, 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 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 two

[29] 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).

[30] In typical charter lotteries, an offer is extended to all siblings when multiple siblings enter the same lottery and one sibling wins.

years after the random lotteries, students who won the lotteries scored, if anything, lower on standardized test scores than students who lost the lotteries.[31] 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 0.07σ to 0.94σ – 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 0.4σ higher per year in math and 0.2σ higher per year in reading (Abdulkadiroglu et al. 2011). Promise Academy students in the Harlem Children's Zone (HCZ) score 0.229σ higher per year in math and 0.047σ 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 0.180σ higher per year in math and 0.075σ higher per year in reading (Tuttle et al. 2013; Angrist et al. 2011). The SEED urban boarding school in Washington D.C., 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 0.16σ with statistically 0 effect on reading.

According to the National Alliance for Public Charter Schools, the median grade served by charter schools in the U.S. is sixth 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 second 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.

[31] The two-year ITT impact for the pooled sample is -0.08σ (p-value = 0.032) for reading scores and -0.06σ (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 0.16σ impact on math scores while schools outside of large urban areas had a -0.14σ impact. Both of these impacts were statistically significant.

Another leading theory posits that reading scores are influenced by the language spoken when students are outside of the classroom (Rickford 1999; Charity, Scarborough, and Griffin 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 non-experimental. 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.

Dobbie and Fryer (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 were 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.

Dobbie and Fryer (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, Dobbie and Fryer (2013) analyzes charter schools in NYC, whereas Krueger (1999) uses a sample of traditional public schools in Tennessee. Second, the variation in

Dobbie and Fryer (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 Dobbie and Fryer (2013) used third through eighth grade test scores. Fourth, and most important, the analysis in Krueger (1999) is experimental.

In stark contrast, Dobbie and Fryer (2013) demonstrate that an index of five policies suggested by forty 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 fifteen 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 $9,959 and $9,448 are, respectively, Austria and Denmark (Aud et al. 2011). However, Austrian fifteen year-olds rank eighteenth in math achievement, twenty-fourth in science, and thirty-first in reading and Danish fifteen 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.[32]

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.

[32] 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).

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 (2014).

Between the 2010-2011 and 2012-2013 school years, Fryer (2014) implemented important elements of the above education best-practices in twenty 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 one hour and the school year was lengthened by ten days in the nine secondary (middle and high) schools. This was 21 percent more time in school than students in these schools spent in the pre-treatment 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 eleven elementary schools, the length of the day and the year were not changed, but non-instructional activities (e.g. twenty-minute bathroom breaks) were reduced. This is consistent with the correlations in Dobbie and Fryer (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, nineteen out of twenty principals were removed and 46 percent 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 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 quasi-experimental program evaluated in Taylor and Tyler (2012).

To enhance student-level differentiation, all fourth, sixth and ninth 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 Dobbie and Fryer (2013).

In order to help teachers use interim data on student performance to guide and inform instructional practice, schools were required to administer interim assessments every three to four 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 Dobbie and Fryer (2013) demonstrate,

data alone is not enough. Dobbie and Fryer (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 eleven of the twenty 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 twenty schools and provide a way to choose between alternative quasi-experimental specifications, we randomly selected nine additional elementary schools (vis-à-vis matched-pairs) from eighteen 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 sixteen schools.

In the sample of sixteen 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. sixth and ninth) or having attended a treatment school in the pre-treatment 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 0.184σ 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 0.146σ 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 Dobbie and Fryer (2013) – fourth, sixth, and ninth grade math – the increase in student achievement is substantially larger than the increase in other grades. Relative to students who attended control schools, fourth graders in treatment schools scored 0.331σ (0.104) higher in math, per year. Similarly, sixth and ninth grade math scores increased 0.608σ (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 like 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).[33] 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 0.4σ 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.

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 non-cognitive 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

$$EC\ Outcome_i = \beta_0^{ec} + \beta_{cab}^{ec} CAB_i + \varepsilon_i^{ec}.$$

Similarly, each of the MC outcomes is given by

$$MC\ Outcome_i = \beta_0^{mc} + \beta_{cab}^{mc} CAB_i + \beta_{ec}^{mc} EC_i + \varepsilon_i^{mc}.$$

For the adolescent life-stage we have

$$AD\ Outcome_i = \beta_0^{ad} + \beta_{cab}^{ad} CAB_i + \beta_{ec}^{ad} EC_i + \beta_{mc}^{ad} MC_i + \varepsilon_i^{ad}.$$

Outcomes when transitioning to adulthood would be

$$TTA\ Outcome_i = \beta_0^{tta} + \beta_{cab}^{tta} CAB_i + \beta_{ec}^{tta} EC_i + \beta_{mc}^{tta} MC_i + \beta_{ad}^{tta} AD_i + \varepsilon_i^{tta}.$$

[33] 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.

And finally, adult outcomes are modeled as

$$AH\ Outcome_i = \beta_0^{ah} + \beta_{eab}^{ah} CAB_i + \beta_{ec}^{ah} EC_i + \beta_{mc}^{ah} MC_i + \beta_{ad}^{ah} AD_i + \beta_{tta}^{ah} TTA_i + \varepsilon_i^{ah}.$$

Where β_ψ^λ are the partial correlations of realized outcomes from the ψ life-stage ("0" represents an intercept) with the given LHS outcome in the λ 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, in order 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), non-cognitive 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. In order 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:

$$TTA\ Outcome_i = \beta_0^{tta} + \beta_{link}^{tta} LINK_i + \varepsilon_i^{tta}$$

$$AH\ Outcome_i = \beta_0^{ah} + \beta_{link}^{ah} LINK_i + \beta_{tta}^{ah} TTA_i + \varepsilon_i^{ah}.$$

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 post-intervention 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.

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 1.192σ in math and 2.449σ in reading.[34] [35] 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.[36] 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 won't 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.

[34] 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.

[35] 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.

[36] Using the cross-sectional estimates generated by Chetty et al. (2014), the expected income gain at age 28 from a 1.192σ increase in standardized math scores is 15.62% and from a 2.449σ increase in standardized reading scores is 32.08%.

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 have less demonstrated success.

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?

References

[1] Aaronson, Daniel. 1998. "Using Sibling Data to Estimate the Impact of Neighborhoods on Children's Educational Outcomes." *Journal of Human Resources*, 33(4): 915-946.

[2] Aaronson, Daniel, Lisa Barrow, and William Sander. 2007. "Teachers and Student Achievement in the Chicago Public High Schools." *Journal of Labor Economics*, 25(1): 95-135.

[3] Abdulkadiroglu, Atila, Joshua Angrist, Susan Dynarski, Thomas Kane, and Parag Pathak. 2011. "Accountability and Flexibility in Public Schools: Evidence from Boston's Charters and Pilots." *The Quarterly Journal of Economics*, 126(2): 699-748.

[4] Administration for Children and Families. 2006. "Preliminary Findings from the Early Head Start Prekindergarten Followup." Washington, D.C.: U.S. Department of Health and Human Services Report.

[5] Ainsworth, James. 2002. "Why Does It Take a Village? The Mediation of Neighborhood Effects on Educational Achievement." *Social Forces*, 81(1): 117-152.

[6] Akerlof, George. 1978. "The Economics of "Tagging" as Applied to the Optimal Income Tax, Welfare Programs, and Manpower Planning." *The American Economic Review*, 68(1): 8-19.

[7] Alexander, Karl, Doris Entwisle, and Linda Olson. 2001. "Schools, Achievement, and Inequality: A Seasonal Perspective." *Educational Evaluation and Policy Analysis*, 23(2): 171-191.

[8] Allington, Richard, Anne McGill-Franzen, Gregory Camilli, Lunetta Williams, Jennifer Graf, Jacqueline Zeig, Courtney Zmach, and Rhonda Nowak. 2010. "Addressing Summer Reading Setback Among Economically Disadvantaged Elementary Students." *Reading Psychology*, 31(5): 411-427.

[9] Almond, Douglas, and Janet Currie. 2011. "Killing Me Softly: The Fetal Origins Hypothesis." *The Journal of Economic Perspectives*, 25(3): 153-172.

[10] Angrist, Joshua, Eric Bettinger, Erik Bloom, Elizabeth King, and Michael Kremer. 2002. "Vouchers for Private Schooling in Colombia: Evidence from a Randomized Natural Experiment." *American Economic Review*, 92(5): 1535-1558.

[11] Angrist, Joshua, Eric Bettinger, and Michael Kremer. 2006. "Long-Term Educational Consequences of Secondary School Vouchers: Evidence from Administrative Records in Colombia." *The American Economic Association*, 96(3): 847-862.

[12] Angrist, Joshua, Susan Dynarski, Thomas Kane, Parag Pathak, and Christopher Walters. 2011. "Who Benefits From KIPP?" IZA Discussion Paper no. 5690.

[13] Angrist, Joshua, Daniel Lang, and Philip Oreopoulos. 2009. "Incentives and Services for College Achievement: Evidence from a Randomized Trial." *American Economic Journal: Applied Economics*, 1(1): 136-163.

[14] Angrist, Joshua, and Victor Lavy. 2009. "The Effects of High Stakes High School Achievement Awards: Evidence from a Randomized Trial." *American Economic Review*, 99(4): 1384-1414.

[15] Angrist, Joshua, Parag Pathak, and Christopher Walters. 2013. "Explaining Charter School Effectiveness." *American Economic Journal: Applied Economics*, 5(4): 1-27.

[16] Aronson, Joshua, Carrie Fried, and Catherine Good. 2002. "Reducing the Effects of Stereotype Threat on African American College Students by Shaping Theories of Intelligence." *Journal of Experimental Social Psychology*, 38: 113-125.

[17] Attewell, Paul, and Battle, Juan. 1999. "Home Computers and School Performance." *The Information Society*, 15: 1-10.

[18] Aud, Susan, William Hussar, Grace Kena, Kevin Bianco, Lauren Frohlich, Jana Kemp, and Kim Tahan. 2011. "The Condition of Education 2011." Washington, D.C.: U.S. Department of Education, National Center for Education Statistics.

[19] Avvisati, Francesco, Marc Gurgand, Nina Guyon, and Eric Maurin. 2014. "Getting Parents Involved: A Field Experiment in Deprived Schools." *Review of Economic Studies*, 81(1): 57-83.

[20] Baker, George. 2002. "Distortion and Risk in Optimal Incentive Contracts." *Journal of Human Resources*, 37(4): 728-751.

[21] Barlevy, Gadi, and Derek Neal. 2012. "Pay for Percentile." *American Economic Review*, 102(5): 1805-1831.

[22] Barrera-Osorio, Felipe, Marianne Bertrand, Leigh Linden, and Francisco Perez-Calle. 2011. "Improving the Design of Conditional Transfer Programs: Evidence from a Randomized Education Experiment in Colombia." *The American Economic Journal: Applied Economics*, 3(2): 167-195.

[23] Becker, Gary. 1995. "Human Capital and Poverty Alleviation." Human Resource and Operations Policy, World Bank, Working Paper no. 52.

[24] Behrman, Jere, Piyali Sengupta, and Petra Todd. 2001. "Progressing Through PROGRESA: An Impact Assessment of a School Subsidy Experiment." Washington, D.C.: International Food Policy Research Institute.

[25] Behrman, Jere, Piyali Sengupta, and Petra Todd. 2005. "Progressing through PROGRESA: An Impact Assessment of a School Subsidy Experiment in Rural Mexico." *Economic Development and Cultural Change*, 54(1): 237-275.

[26] Berger, Andrea, Turk-Bicakci, Lori, Garet, Michael, Song, Mengli, Knudson, Joel, Haxton, Clarisse, Zeiser, Kristina, Hoshen, Gur, Ford, Jennifer, Stephan, Jennifer. 2013. "Early College, Early Success:

Early College High School Initiative Impact Study." Washington, D.C.: American Institutes for Research

[27] Bettinger, Eric. 2012. "Paying to Learn: The Effect of Financial Incentives on Elementary School Test Scores." *The Review of Economics and Statistics*, 94(3): 686-698.

[28] Bifulco, Robert, Casey Cobb, and Courtney Bell. 2009. "Can Interdistrict Choice Boost Student Achievement? The Case of Connecticut's Interdistrict Magnet School Program." *Educational Evaluation and Policy Analysis*, 31(4): 323-345.

[29] Bill and Melinda Gates Foundation. 2014. "Teacher's Know Best: Teachers' Views on Professional Development." Seattle, WA: Bill and Melinda Gates Foundation.

[30] Blachman, Benita. 1987. "An Alternative Classroom Reading Program for Learning Disabled and Other Low-Achieving Children." In Bowler R., editor. *Intimacy with Language: A Forgotten Basic in Teacher Education*, Baltimore, MD: Orton Dyslexia Society.

[31] Blachman, Benita, Darlene Tangel, Eileen Wynne Ball, Rochella Black, and Collen McGraw. 1999. "Developing Phonological Awareness and Word Recognition Skills: A Two-Year Intervention with Low-Income, Inner-City Children." *Reading and Writing*, 11(3): 239-273.

[32] Blachman, Benita, Christopher Schatschneider, Jack Fletcher, David Francis, Sheila Clonan, Bennett Shaywitz, and Sally Shaywitz. 2004. "Effects of Intensive Reading Remediation for Second and Third Graders and a 1-Year Follow-Up." *Journal of Educational Psychology*, 96(3): 444-461.

[33] Bloom, Nicholas, Renata Lemos, Raffaella Sadun, and John Van Reenen. 2015. "Does Management Matter in Schools?" *The Economic Journal*, 125(584): 647-674.

[34] Borman, Geoffrey, Robert Slavin, Alan Cheung, Anne Chamberlain, Nancy Madden, and Bette Chambers. 2007. "Final Reading Outcomes of the National Randomized Field Trial of Success for All." *American Education Research Journal*, 44(3): 701-731.

[35] Boyd, Donald, Daniel Goldhaber, Hamilton Lanjford, and James Wyckoff. 2007. "The Effect of Certification and Preparation on Teacher Quality." *The Future of Children*, 17(1): 45-68.

[36] Broh, Beckett. 2004. "Racial/Ethnic Achievement Inequality: Separating School and Non-School Effects Through Seasonal Comparisons." Dissertation submitted to Ohio State University, Athens, OH.

[37] Brooks-Gunn, Jeanne, and Greg Duncan. 1997. "The Effects of Poverty on Children." *The Future of Children*, 7(2): 55-71.

[38] Brooks-Gunn, Jeanne, Pamela Klebanov, Judith Smith, Greg Duncan, and Kyunghee Lee. 2003. "The Black-White Test Score Gap in Young Children: Contributions of Test and Family Characteristics." *Applied Developmental Science*, 7(4): 239-252.

[39] Brooks-Gunn, Jeanne, Fong-Ruey Liaw, and Pamela Kato Klebanov. 1992. "Effects of Early Intervention on Cognitive Function of Low Birth Weight Preterm Infants." *The Journal of Pediatrics*, 120(3): 350-359.

[40] Campbell, Frances, and Craig Ramey. 1994. "Effects of Early Intervention on Intellectual and Academic Achievement: A Follow-Up Study of Children from Low-Income Families." *Child Development*, 65(2): 684-698.

[41] Carlson, Deven, Geoffrey Borman, Michelle Robinson. 2011. "A Multistate District-Level Cluster Randomized Trial of the Impact of Data-Driven Reform on Reading and Mathematics Achievement." *Educational Evaluation and Policy Analysis*, 33(3): 378-398.

[42] Carneiro, Pedro, and James Heckman. 2003. "Human Capital Policy." NBER Working Paper no. 9495.

[43] Center for Research on Education Outcomes (CREDO). 2013. "National Charter School Study." Stanford, CA: Center for Research on Education Outcomes.

[44] Center, Yola, Kevin Wheldall, Louella Freeman, Lynne Outhred and Margaret McNaught. 1995. "An Evaluation of Reading Recovery." *Reading Research Quarterly*, 30(2): 240-263.

[45] Charity, Anne, Hollis Scarborough, and Darion Griffin. 2004. "Familiarity with School English in African American Children and Its Relation to Early Reading Achievement." *Child Development*, 75(5): 1340-1356.

[46] Chase-Lansdale, Lindsay, and Rachel Gordon. 1996. "Economic Hardship and the Development of Five-and Six-Year-Olds: Neighborhood and Regional Perspectives." *Child Development*, 67(6): 3338-3367.

[47] Chase-Landsdale, Lindsay, Rachel Gordon, Jeanne Brooks-Gunn, and Pamela K. Klebanov. 1997. "Neighborhood and Family Influences on the Intellectual and Behavioral Competence of Preschool and Early School-Age Children"In: Jeanne Brooks-Gunn, Greg Duncan, and J. Lawrence Aber, editors.

Neighborhood Poverty: Context and Consequences for Children, Volume 1. 79-118. New York: Russel Sage.

[48] Chenoweth, Karin. 2007. ""It's Being Done": Academic Success in Unexpected School." Cambridge, MA: Harvard Education Press.

[49] Chetty, Raj, John Friedman, Nathaniel Hilger, Emmanuel Saez, Diane Whitmore Schanzenbach, and Danny Yagan. 2011. "How Does Your Kindergarten Classroom Affect Your Earnings? Evidence from Project STAR." *The Quarterly Journal of Economics*, 126(4): 1593-1660.

[50] Chetty, Raj, John Friedman, and Jonah Rockoff. 2014. "Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood." *American Economic Review*, 104(9): 2633-2679.

[51] Chetty, Raj, Nathaniel Hendren, and Lawrence Katz. 2016. "The Effects of Exposure to Better Neighborhoods on Children: New Evidence from the Moving to Opportunities Experiment." *American Economic Review*, forthcoming.

[52] Clark, Melissa, Hanley Chiang, Tim Silva, Sheena McConnell, Kathy Sonnenfeld, and Anastasia Erbe. 2013. "The Effectiveness of Secondary Math Teachers from Teach for America and the Teaching Fellows Programs." Washington, D.C.: National Center for Educational Evaluation and Regional Assistance.

[53] Cohen, David, and Heather Hill. 2001. "Learning Policy: When State Education Reform Works." New Haven, CT: Yale University Press.

[54] Cohen, Geoffrey, Julio Garcia, Nancy Apfel, Allison Master. 2006. "Reducing the Racial Achievement Gap: A Social-Psychological Intervention." *Science*, 313(5791): 1307-1310.

[55] Cohen, Geoffrey, Julio Garcia, Valerie Purdie-Vaughns, Nancy Apfel, Patricia Brzustoski. 2009. "Recursive Processes in Self-Affirmation: Intervening to Close the Minority Achievement Gap." *Science*, 324(5925): 400-403.

[56] Coleman, James, Ernest Campbell, Carol Hobson, James McPartland, Alexander Mood, Frederic Weinfeld, and Robert York. 1966. "Equality of Educational Opportunity." Washington, D.C.: U.S. Department of Health, Education, and Welfare.

[57] Constantine, Jill, Daniel Player, Tim Silva, Kristin Hallgren, Mary Grider, and John Deke. 2009. "An Evaluation of Teachers Trained Through Different Routes to Certification." Washington, D.C.: National Center for Education Evaluation and Regional Assistance.

[58] Cook, Philip, Kenneth Dodge, George Farkas, Roland Fryer, Jonathan Guryan, Jens Ludwig, Susan Mayer, Harold Pollack, and Laurence Steinberg. 2014. "The (Surprising) Efficacy of Academic and Behavioral Intervention with Disadvantaged Youth: Results from a Randomized Experiment in Chicago." NBER Working Paper no. 19862.

[59] Cooper, Harris, Barbara Nye, James Lindsay, and Scott Greathouse. 1996. "The Effects of Summer Vacation on Achievement Test Scores: A Narrative and Meta-Analytic Review." *Review of Educational Research*, 66(3): 227-268.

[60] Corcoran, Sean, William Evans, and Robert Schwab. 2004. "Changing Labor-Market Opportunities for Women and the Quality of Teachers, 1957-2000." *American Economic Review*, 94(2): 729-760.

[61] Courtney, Mark, Andrew Zinn, Erica Zielewski, Roseana Bess, and Karin Malm. 2008. "Evaluation of the Early Start to Emancipation Preparation—Tutoring Program Los Angeles County, California: Final Report." Washington, D.C.: The Urban Institute.

[62] Cullen, Julie Berry, Brian Jacob, and Steven Levitt. 2006. "The Effect of School Choice on Participants: Evidence from Randomized Lotteries." *Econometrica*, 74(5): 1191-1230.

[63] Cullen, Julie Berry, Steven Levitt, Erin Robertson, and Sally Sadoff. 2013. "What Can be Done to Improve Struggling High Schools?" *The Journal of Economic Perspectives*, 27(2): 133-152.

[64] Cunha, Flavio, and James Heckman. 2007. "The Technology of Skill Formation." *The American Economic Review*, 97(2): 31-47.

[65] Cunha, Flavio, and James Heckman. 2010. "Investing in Our Young People" NBER Working Paper no. 16201.

[66] Currie, Janet, and Duncan Thomas. 1995. "Does Head Start Make a Difference." *The American Economic Review*, 85(3): 341-364.

[67] Curto, Vilsa, and Roland Fryer. 2014. "The Potential of Urban Boarding Schools for the Poor." *Journal of Labor Economics*, 32(1): 65-93.

[68] Davis-Kean, Pamela. 2005. "The Influence of Parent Education and Family Income on Child Achievement: The Indirect Role of Parental Expectations and the Home Environment." *Journal of Family Psychology* 19(2): 294-304.

[69] de la Rica, Sara. 2011. "Social and Labor Market Integration of Ethnic Minorities in Spain." In: Martin Kahanec, and Klaus Zimmerman, editors. *Ethnic Diversity in European Labor markets: Challenges and Solutions*. 268-282. Cheltenham, UK: Edward Elgar Publishing.

[70] Deaton, Angus. 2010. "Instruments, Randomization, and Learning about Development." *The Journal of Economic Literature*, 48(2): 424-455.

[71] Dee, Thomas, and Brian Jacob. 2011. "The Impact of No Child Left Behind on Student Achievement." *Journal of Policy Analysis and Management*, 30(3): 418-446.

[72] Deming, David, Sarah Cohodes, Jennifer Jennings, and Christopher Jencks. 2013. "School Accountability, Postsecondary Attainment and Earnings." NBER Working Paper no. 19444.

[73] DerSimonian, Rebecca, and Nan Laird. 1986. "Meta-analysis in Clinical Trials." *Control Clin Trials*, 7(3): 177-188.

[74] Dobbie, Will, and Roland Fryer. 2011. "Are High-Quality Schools Enough to Increase Achievement Among the Poor? Evidence from the Harlem Children's Zone." *American Economic Journal: Applied Economics*, 3(3): 158-187.

[75] Donaldson, Morgaen, and Susan Moore Johnson. 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." *Educational Evaluation and Policy Analysis*, 32(2): 299-323.

[76] Duckworth, Angela, Teri Kirby, Anton Gollwitzer, and Gabrielle Oettingen. 2013. "From Fantasy to Action: Mental Contrasting With Implementation Intentions (MCII) Improves Academic Performance in Children." *Social Psychological and Personality Science*, 4: 745-753.

[77] Duckworth, Angela, and Martin Seligman. 2005. "Self-Discipline Outdoes IQ in Predicting Academic Performance of Adolescents." *Psychological Science*, 16(12): 939-944.

[78] Duflo, Esther, Rema Hanna, and Stephne Ryan. 2012. "Incentives Work: Getting Teachers to Come to School." *American Economic Review*, 102(4): 1241-1278.

[79] Duncan, Greg, Jeanne Brooks-Gunn, and Pamela Kato Klebanov. 1994. "Economic Deprivation and Early Childhood Development." *Child Development*, 65(2): 296-318.

[80] Duncan, Greg, and Katherine Magnuson. 2005. "Can Family Socioeconomic Resources Account for Racial and Ethnic Test Score Gaps?" *Future of Children*, 15(1): 35-54.

[81] Dunstan, William. 2010. "Ancient Rome." Lanham, MD: Rowman and Littlefield.

[82] Evans, Gary. 2004. "The Environment of Childhood Poverty." *American Psychologist*, 59(2): 77-92.

[83] Fairlie, Robert. 2005. "The Effects of Home Computers on School Enrollment." *Economics of Education Review*, 24(5): 533-547.

[84] Fairlie, Robert, Daniel Beltran, and Kuntal Das. 2010. "Home Computers and Educational Outcomes: Evidence from the NLSY97 and CPS." *Economic Inquiry*, 48(3): 771-792.

[85] Fairlie, Robert, and Jonathan Robinson. 2013. "Experimental Evidence on the Effects of Home Computers on Academic Achievement among Schoolchildren." *American Economic Journal: Applied Economics*, 5(3): 211-240.

[86] Fantuzzo, John, Gwendolyn Davis, and Marika Ginsburg. 1995. "Effects of Parent Involvement in Isolation or in Combination with Peer Tutoring on Student Self-Concept and Mathematics Achievement." *Journal of Educational Psychology*, 87(2): 272-281.

[87] Feng, Li. 2010. "Hire Today, Gone Tomorrow: New Teacher Classroom Assignments and Teacher Mobility." *Educational Finance and Policy*, 5(3): 278-316.

[88] Fiorini, Mario. 2010. "The effect of home computer use on children's cognitive and non-cognitive skills." *Economics of Education Review*, 29(1): 55-72.

[89] Fleischman, Howard, Paul Hopstock, Marisa Pelczar, and Brooke Shelley. 2010. "Highlights from PISA 2009: Performance of U.S. 15-Year-Old Students in Reading, Mathematics, and Science Literacy in an International Context." Washington, D.C.: U.S. Department of Education.

[90] Fletcher, Jack, and Reid Lyon. 1998. "Reading: A Research-Based Approach"in "What's Gone Wrong in America's Classrooms"Stanford, CA: Hoover Institution Press.

[91] Foorman, Barbara, and Moats, Louisa. 2004. "Conditions for Sustaining Research-Based Practices in Early Reading Instruction." *Remedial Special Education*, 25(1): 51-60.

[92] Friedman, Milton. 1955. "The Role of Government in Public Education." In: Robert Solo, editor. *Economics and the Public Interest*. New Brunswick, NJ: University of Rutgers Press.

[93] Fryer, Roland. 2010. "Racial Inequality in the 21st Century: The Declining Significance of Discrimination." *Handbook of Labor Economics*, 4(B): 855-971.

[94] Fryer, Roland. 2011. "Financial Incentives and Student Achievement: Evidence from Trials." *The Quarterly Journal of Economics*, 126(4): 1755-1798.

[95] Fryer, Roland. 2013. "Information and Student Achievement: Evidence from a Cellular Phone Experiment"NBER Working Paper no. 19113.

[96] Fryer, Roland. 2013. "Teacher Incentives and Student Achievement: Evidence from New York City Public Schools." *Journal of Labor Economics*, 31(2): 373-427.

[97] Fryer, Roland. 2014. "Injecting Charter School Nest Practices into Traditional Public Schools: Evidence From Field Experiments." *Quarterly Journal of Economics*, 129(3): 1355-1407.

[98] Fryer, Roland. 2014. "Teacher Incentives and Student Achievement: Evidence from New York City Public Schools." *Quarterly Journal of Economics*, 129(3):1355-1407.

[99] Fryer, Roland, and Will Dobbie. 2013. "Getting Beneath the Veil of Effective Schools: Evidence from New York City." *American Economic Journal: Applied Economics*, 5(4): 28-60.

[100] Fryer, Roland, and Richard Holden. 2013. "Multitasking, Dynamic Complementaries, and Incentives: A Cautionary Tale." Working Paper.

[101] Fryer, Roland, and Steven Levitt. 2004. "Understanding the Black-White Test Score Gap in the First Two Years of School." *The Review of Economic and Statistics*, 86(2): 447-464.

[102] Fryer, Roland, and Steven Levitt. 2006. "The Black-White Test Score Gap Through Third Grade." *American Law and Economic Review*, 8(2): 249-281.

[103] Fryer, Roland, and Steven Levitt. 2013. "Testing for Racial Differences in the Mental Ability of Young Children." *American Economic Review*, 103(2): 981-1005.

[104] Fryer, Roland, Steven Levitt, and John List. 2015. "Parental Incentives and Early Childhood Achievement: A Field Experiment in Chicago Heights." Unpublished working paper.

[105] Fryer, Roland, Steven Levitt, John List, and Sally Sadoff. 2015. "Enhancing the Efficacy of Teacher Incentives Through Loss Aversion: A Field Experiment." NBER Working Paper no. 18237.

[106] Fuchs, Thomas, and Ludger Woessmann. 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.

[107] Garber, Howard. 1988. "The Milwaukee Project: Preventing Mental Retardation in Children at Risk." Washington, D.C.: National Institute of Handicapped Research.

[108] Garces, Eliana, Duncan Thomas, and Janet Currie. 2002. "Longer-Term Effects of Head Start." *The American Economic Review*, 92(4): 999-1012.

[109] Garet, Michael, Stephanie Cronen, Marian Eaton, Anja Kurki, Wehmah Jones, Kazuaki Uekawa, and Audrey Falk. 2008. "The Impact of Two Professional Development Interventions on Early Reading Instruction and Achievement." Washington, D.C.: National Center for Education Evaluation and Regional Assistance.

[110] Garet, Michael, Andrew Porter, Laura Desimone, Beatrice Birman, Kwang Suk Yoon. 2001. "What Makes Professional Development Effective? Results from a National Sample of Teachers." *American Educational Research Journal*, 38(4): 915-945.

[111] Glass, Gene, and Mary Lee Smith. 1978. "Meta-Analysis of Research on The Relationship of Class-Size and Achievement." San Francisco, CA: Far West Laboratory for Educational Research and Development.

[112] Glazerman, Steven, Daniel Mayer, and Paul Decker. 2006. "Alternative Routes to Teaching: The Impacts of Teach for America on Student Achievement and Other Outcomes." *Journal of Policy Analysis and Management*, 25(1): 75-96.

[113] Glazerman, Steven, Allixon McKie, and Nancy Carey. 2009. "An Evaluation of the Teacher Advancement Program (TAP) in Chicago: Year One Impact Report." Princeton, NJ: Mathematica Policy Research.

[114] Glazerman, Steven, Ali Protik, Bing-ru The, Julie Bruch, Jeffrey Max. 2013. "Transfer Incentives for High-Performing Teachers: Final Results from a Multisite Randomized Experiment." Washington, D.C.: National Center for Education Evaluation and Regional Assistance.

[115] Gleason, Phillip, Melissa Clark, Christina Tuttle, and Emily Dwoyer. 2010. "The Evaluation of Charter School Impacts." Washington, D.C.: National Center for Education Evaluation and Regional Assistance.

[116] Glewwe, Paul, Nauman Ilias, and Michael Kremer. 2010. "Teacher Incentives." *American Economic Journal: Applied Economics*, 2(3): 205-227.

[117] Gray, Susan, and Rupert Klaus. 1970. "The Early Training Project: A Seventh-Year Report." *Child Development*, 41: 909-924.

[118] Greene, Jay, Paul Peterson, and Jiangtao Du. 1999. "Effectiveness of School Choice: The Milwaukee Experiment." *Education and Urban Society*, 31(2) : 190-213.

[119] Hahn, Andrew, Tom Leavitt, and Paul Aaron. 1994. "Evaluation of the Quantum Opportunities Program (QOP) Did the Program Work?" Waltham, MA: Brandeis University Center for Human Resources.

[120] Halpern-Felsher, Bonnie, James P. Connell, Margaret Beale Spencer, J. Lawrence Aber, Greg P. Duncan, Elizabeth Clifford, Warren E. Crichlow, Peter A. Usinger, Steven P. Cole, LaRue Allen, and Edward Seidman. 1997. "Neighborhood and Family Factors Predicting Educational Risk and Attainment in African American and White Children and Adolescents." In: Jeanne Brooks-Gunn, Greg Duncan, and Lawrence Aber, editors. *Neighborhood Poverty, Volume I: Context and Consequences for Children*. New York: Russel Sage Foundation.

[121] Hamilton, Gayle, Stephen Freedman, Lisa Gennetian, Charles Michalopoulos, Johanna Walter, Diana Adams-Ciardullo, Anna Gassman-Pines. 2001. "National Evaluation of Welfare-to-Work Strategies." Washington, D.C.: U.S. Department of Health and Human Services.

[122] Hanushek, Eric. 1979. "Conceptual and Empirical Issues in the Estimation of Educational Production Functions." *The Journal of Human Resources*, 14(3): 351-388.

[123] Harrington, Michael. 1982. "The Other America: Poverty in the United States." New York, NY: Touchstone.

[124] Harrison, Glenn, and John List. 2004. "Field Experiments." *The Journal of Economic Literature*, 42(4): 1009-1055.

[125] Hatton, Timothy. 2011. "The Social and Labor Market Outcomes of Ethnic Minorities in the UK." In: Martin Kahanec, and Klaus Zimmerman, editors. *Ethnic Diversity in European Labor markets: Challenges and Solutions*. 283-306. Cheltenham, UK: Edward Elgar Publishing.

[126] Heckman, James. 2008. "Role of Income and Family Influence on Child Outcomes." *Annals of the New York Academy of Sciences*, 1136: 307-323.

[127] Heckman, James, Seong Hyeok Moon, Rodrigo Pinto, Peter Savelyev, and Adam Yavitz. 2009. "A Reanalysis of the High/Scope Perry Preschool Program." University of Chicago, Department of Economics, Unpublished Manuscript.

[128] Heckman, James, Seong Hyeok Moon, Rodrigo Pinto, Peter Savelyev, and Adam Yavitz. 2010. "The Rate of Return to the High/Scope Perry Preschool Program." *Journal of Public Economics*, 94(1-2): 114-128.

[129] Heckman, James, and Yona Rubenstein. 2001. "The Importance of Noncognitive Skills: Lessons from the GED Testing Program." *The American Economic Review*, 91(2): 145-149.

[130] Heckman, James, Jora Stixrud, and Sergio Urzua. 2006. "The Effects of Cognitive and Noncognitive Abilities on Labor Market Outcomes and Social Behavior." *Journal of Labor Economics*, 24(3): 411-482.

[131] Heckman, James, and Tim Kautz. 2013. "Fostering and Measuring Skills: Interventions That Improve Character and Cognition." NBER Working Paper no. 19656.

[132] Hedges, Larry. 1981. "Distribution Theory for Glass's Estimator of Effect Sizes and Related Estimators." *Journal of Educational and Behavioral Statistics*, 6(2): 107-128.

[133] Heyns, Barbara. 1978. "Summer Learning and the Effects of Schooling." Orlando, FL: Academic Press.

[134] Heyns, Barbara. 1987. "Schooling and Cognitive Development." *Child Development*, 58(5): 1151-1160.

[135] Hill, Hillary. 2007. "Learning in the Teaching Workforce." *The Future of Children*, 17(1): 111-127.

[136] Hirst, Lois Trimble. 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.

[137] Holmes, Thomas, and Barbara McConnell. 1990. "Full-Day Versus Half-Day Kindergarten: An Experimental Study." Paper presented at the annual meeting of the Educational Research Association, Boston, MA.

[138] Holmstrom, Bengt, and Paul Milgrom. 1991. "Multitask Principal-Agent Analyses: Incentive Contracts, Asset Ownership, and Job Design." *Journal of Law, Economics, & Organization*, 7: 24-52.

[139] Hopkins, Kenneth, and Glenn Bracht. 1975. "Ten-Year Stability of Verbal and Nonverbal IQ Scores." *American Education Research Journal*, 12(4): 469-477.

[140] Hoxby, Caroline. 2002. "School Choice and School Productivity." NBER Working Paper no. 8873.

[141] Hoxby, Caroline, and Andrew Leigh. 2004. "Pulled Away or Pushed Out? Explaining the Decline of Teacher Aptitude in the United States." *American Economic Review*, 94(2): 236-240.

[142] Hoxby, Caroline, and Sonali Murarka. 2009. "Charter Schools in New York City: Who Enrolls and How They Affect Their Students' Achievement." NBER Working Paper no. 14852.

[143] Jackson, Clement. 2010. "A Stitch in Time: The Effects of a Novel Incentive-Based High-School Intervention on College Outcomes." NBER Working Paper no. w15722.

[144] Jacob, Brian. 2004. "Remedial Education and Student Achievement: A Regression-Discontinuity Analysis." *The Review of Economics and Statistics*, 86(1): 226-244.

[145] Jeynes, William. 2005. "Parental Involvement and Student Achievement: A Meta-Analysis." Cambridge, MA: Harvard Family Research Project.

[146] Jeynes, William. 2007. "The Relationship Between Parental Involvement and Urban Secondary School Student Academic Achievement: A Meta-Analysis." *Urban Education*, 42(1): 82-110.

[147] Joyce, Bruce and Beverly Showers. 1988. "Student Achievement Through Staff Development." White Plains, NY: Longman.

[148] Kántor, Zoltán. 2011. "Ethnic or Social Integration? The Roma in Hungary." In: Martin Kahanec, and Klaus Zimmerman, editors. *Ethnic Diversity in European Labor markets: Challenges and Solutions*. 137-162. Cheltenham, UK: Edward Elgar Publishing.

[149] Kane, Thomas, and Douglas Staiger. 2008. "Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation." NBER Working Paper no. 14607.

[150] Kautz, Tim, James Heckman, Ron Diris, Bas ter Weel, and Lex Borghans. 2014. "Fostering and Measuring Skills: Improving Cognitive and Non-Cognitive Skills to Promote Lifetime Success." NBER Working Paper no. 20749.

[151] Kim, James. 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.

[152] Klibanoff, Leonard, and Sue Haggart. "Summer Growth and the Effectiveness of Summer School." Mountainview, CA: RMC Research Corporation.

[153] Kling, Jeffrey, Jeffrey Liebman, and Lawrence Katz. 2007. "Experimental Analysis of Neighborhood Effects." *Econometrica*, 75(1): 83-119.

[154] Knudson, Eric, James Heckman, Judy Cameron, and Jack Shonkoff. "Economic, Neurobiological, and Behavioral Perspectives on Building America's Future Workforce." *Proceedings of the National Academy of Sciences*, 103(27): 10155-10162.

[155] Kohen, Dafna, Jeanne Brooks-Gunn, Tama Leventhal, and Clyde Hertzman. 2002. "Neighborhood Income and Physical and Social Disorder in Canada: Associations with Young Children's Competencies." *Child Development* 73(6): 1844-1860.

[156] Kremer, Michael, Edward Miguel, Rebecca Thornton. 2009. "Incentives to Learn." *The Review of Economics and Statistics*, 91(3): 437-456.

[157] Krueger, Alan. 1999. "Experimental Estimates of Education Production Functions." *The Quarterly Journal of Economics*, 114(2): 497-532.

[158] Layton, Lyndsey. 2015. "Study: Billions of Dollars in Annual Teacher Training is Largely A Waste." *The Washington Post*. August 04, 2015.

[159] Layzer, Jean, Carolyn Layzer, Barbara Goodson, and Cristofer Price. 2007. "Evaluation of Child Care Subsidy Strategies: Findings From Project Upgrade in Miami-Dade County." Cambridge, MA: Abt Associates.

[160] Levitt, Steven, and John List. 2009. "Field experiments in economics: The past, the present, and the future." *The European Economic Review*, 53(1): 1-18.

[161] Lipsey, Mark, and David Wilson. 2000. "Practical Meta Analysis." Thousand Oaks, California: Sage Publications.

[162] Loucks-Horsley, Susan, Susan Mundry, Peter Hewson, Nancy Love, and Katherine Stiles. 1998. "Designing Professional Development for Teachers of Mathematics and Science." Thousand Oaks, CA: Corwin Press.

[163] Ludwig, Jens, Greg Duncan, Lisa Gennetian, Lawrence Katz, Ronald Kessler, Jeffrey Kling, and Lisa Sanbonmatsu. 2012. "Neighborhood Effects on the Long-Term Well-Being of Low-Income Adults." *Science*, 337(6101): 1505-1510.

[164] Ludwig, Jens, Lisa Sanbonmatsu, Lisa Gennetian, Emma Adam, Greg Duncan, Lawrence Katz, Ronald Kessler, Jeffrey Kling, Stacy Tessler Lindau, Robert Whitaker, and Thomas McDade. 2011. "Neighborhoods, Obesity, and Diabetes—A Randomized Social Experiment." *The New England Journal of Medicine*, 365(16): 1509-1519.

[165] Magnuson, Katherine and Greg Duncan. 2002. "Parents in Poverty." In: Marc Bornstein, editor. *Handbook of Parenting: Second Edition*. Mahwah, NJ: Lawrence Erlbaum Associates.

[166] Malamud, Ofer, and Cristian Pop-Eleches. 2011. "Home Computer Use and the Development of Human Capital." *The Quarterly Journal of Economics*, 126(2): 987-1027.

[167] Marshall, Helen, and Lucille Magruder. 1960. "Relations between Parent Money Education Practices and Children's Knowledge and Use of Money." *Child Development*, 31(2): 253-284.

[168] Mathes, Patricia, and Allison Babyak. 2001. "The Effects of Peer-Assisted Learning Strategies for First-Grade Readers With and Without Additional Mini-Skills Lessons." *Learning Disabilities Research and Practice*, 16(1): 28-44.

[169] May, Henry, Abigail Gray, Jessica Gillespie, Philip Sirinides, Cecile Sam, Heather Goldsworthy, Michael Armijo, and Manrata Tognatta. 2013. "Evaluation of the i3 Scale-up of Reading Recovery." Philadelphia, PA: CPRE.

[170] Mayer, Daniel, Paul Peterson, David Myers, Christina Clark Tuttle, William Howell. 2002. "School Choice in New York City After Three Years: An Evaluation of the School Choice Scholarships Program. Final Report." Princeton, NJ: Mathematica Policy Research.

[171] Maynard, Rebecca, and Richard Murnane. 1979. "The Effects of a Negative Income Tax on School Performance: Results of an Experiment." *The Journal of Human Resources*, 14(4): 463-476.

[172] McCall, William. 1923. "How to Experiment in Education." New York: Macmillan.

[173] McLoyd, Vonnie. 1998. "Socioeconomic Disadvantage and Child Development." *American Psychologist*, 53(2): 185-204.

[174] Meyer, Erik, and Chris Van Klaveren. 2013. "The Effectiveness of Extended Day Programs: Evidence from a Randomized Field Experiment in the Netherlands." *Economics of Education Review*, 36(C): 1-11.

[175] Meyers, Coby, Ayrin Molefe, Sonica Dhillion, and Bo Zhu. 2015. "The Impact of eMINTS Professional Development on Teacher Instruction and Student Achievement." Washington, D.C.: American Institutes for Research.

[176] Michalopoulous, Charles, Dough Tattrie, Cynthia Miller, Philip K. Robins, Pamela Morris, David Gyarmati, Cindy Redcross, Kelly Foley, and Rueben Ford. 2002. "Making Work Pay: Final Report on the Self-Sufficiency Project for Long-Term Welfare Recipients." Ottowa, Canada: Social Research and Demonstration Corporation.

[177] Mischel, Walter, Ebbe Ebbesen, Antonette Raskoff Zeiss. 1972. "Cognitive and Attentional Mechanisms in Delay of Gratification." *Journal of Personality and Social Psychology*, 21(2): 204-218.

[178] Miyake, Akira, Lauren Kost-Smith, Noah Finkelstein, Steven Pollock, Geoffrey Cohen, and Tiffany Ito. 2010. "Reducing the Gender Achievement Gap in College Science: A Classroom Study of Values Affirmation." *Science*, 330(60008): 1234-1237.

[179] Morais de Sá e Silva, Michelle. 2008. "Opportunity NYC: A Performance-based Conditional Cash Transfer Programme." International Poverty Centre Working Paper no. 49.

[180] Mosteller, Frederick, and Robert Boruch. 2002. "Evidence Matters: Randomized Trials in Education Research." Washington, D.C.: Brookings Institution Press.

[181] Moynihan, Daniel. 1969. "On Understanding Poverty: Perspectives from the Social Sciences." New York, NY: Basic Books.

[182] Muralidharan, Karthik, and Venkatesh Sundararaman. 2011. "Teacher Performance Pay: Experimental Evidence from India." *Journal of Political Economy*, 119(1): 39-77.

[183] Murnane, Richard. 1975. "The Impact of School Resources on the Learning of Inner City Children." Cambridge, MA: Ballinger Publishing.

[184] National Alliance for Public Charter Schools. 2009. Public Charter Schools Dashboard, Charter School Market Share.

[185] National Telecommunication and Information Administration. 2011. "Exploring the Digital Nation: Computer and Internet Use at Home." Washington, D.C.: National Telecommunications and Information Administration, U.S. Department of Commerce.

[186] Neal, Derek and William Johnson. 1996. "The Role of Premarket Factors in Black-White Wage Differences." *The Journal of Political Economy*, 104(5): 869-895.

[187] Neal, Derek. 2011. "The Design of Performance Pay Systems in Education." NBER Working Paper no. 16710.

[188] Nelson, Charles. 2000. "Neural Plasticity and Human Development: The Role of Early Experience in Sculpting Memory Systems." *Developmental Science*, 3(2): 115-136.

[189] Nelson, Richard. 1959. "An Experiment with Class Size in the Teaching of Elementary Economics." *Educational Record*, 4: 330-241.

[190] Nelson-Royes, Andrea. 2015. "Why Tutoring?: A Way to Achieve Success in School." Lanham, MD: Rowman and Littlefield.

[191] Newport, Elissa. 1990. "Maturational Constraints on Language Learning." *Cognitive Science*, 14(11): 11-28.

[192] Nordin, Martin and Dan-Olof Rooth. 2007. "Income Gap Between Natives and Second Generation Immigrants in Sweden: Is Skill the Explanation?" IZA Discussion Paper no. 2759.

[193] Nye, Chad, Herb Turner, and Jamie Schwartz. 2006. "Approaches to Parental Involvement for Improving the Academic Performance of Elementary School Children in Grades K-6." Cambridge, MA: Harvard Family Research Project.

[194] O'Neill, June. 1990. "The Role of Human Capital in Earnings Differences Between Black and White Men." *The Journal of Economic Perspectives*, 4(4): 25-45.

[195] Olds, David, Charles Henderson, and Robert Cole. 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.

[196] Olds, David, JoAnn Robinson, and Roth O'Brien. 2002. "Home Visiting by Paraprofessionals and Nurses: A Randomized, Controlled Trial." *Pediatrics*, 100: 486-496.

[197] Oreopoulos, Phillip. 2003. "The Long-Run Consequences of Living in a Poor Neighborhood." *The Quarterly Journal of Economics*, 118(4): 1533-1575.

[198] Parsad, Basmat, Laurie Lewis, and Elizabeth Farris. 2001. "Teacher Preparation and Professional Development: 2000." Washington, D.C.: U.S. Department of Education, National Center for Education Statistics.

[199] Parsons, Craig, and Timothy Smeeding. 2008. "Immigration and the Transformation of Europe." Cambridge, U.K.: Cambridge University Press.

[200] Phillips, Deborah, and Jack Shonkoff. 2000. "From Neurons to Neighborhoods: The Science of Early Childhood Development." Washington, D.C.: National Academies Press.

[201] Phillips, Meredith, Jeanne Brooks-Gun, Greg Duncan, Pamel Klebanov, and Jonathan Crane. 1998. "Family Background, Parenting Practices, and the Black-White Test Score Gap." In: Christopher Jenks, and Meredith Phillips, editors. *The Black-White Test Score Gap*. 102-145. Washington, D.C.: Brookings Institution Press.

[202] Pinkner, Steven. "The Language Instinct." New York, NY: Harper Perennial Modern Classics.

[203] Porwell, P.J. 1978. "Class size: A summary of research." Arlington, VA: Educational Research Service.

[204] Powell-Smith, Kelly, Gary Stoner, Mark Shinn, Roland Good III. 2000. "Parent Tutoring in Reading Using Literature and Curriculum Materials: Impact on Student Reading Achievement." *School Psychology Review*, 29(1): 5-27.

[205] Puma, Michael, Stephen Bell, Ronna Cook, and Camilla Heid. 2010. "Head Start Impact Study Final Report." Washington, D.C.: U.S. Department of Health and Human Services, Administration for Children and Families.

[206] Randel, Bruce, Andrea Beesley, Helen Apthorp, Tedra Clark, Xin Wang, Louis Cicchinelli, and Jean Williams. 2011. "Classroom Assessment for Student Learning: Impact on Elementary School Mathematics in the Central Region." Washington, D.C.: National Center for Educational Evaluation and Regional Assistance.

[207] Riccio, James, Nadine Dechausay, Cynthia Miller, Stephen Nuñez, Nandita Verma, and Edith Yang. 2013. "Conditional Cash Transfers in New York City: The Continuing Story of the Opportunity NYC-Family Rewards Demonstration." New York: MDRC.

[208] Rickford, John. 1999. "African American Vernacular English: Features, Evolution, Educational Implication." Malden, MA: Blackwell Publishers.

[209] Rivkin, Steven, Eric Hanushek, and John Kain. 2005. "Teachers, Schools, and Academic Achievement." *Econometrica*, 73(2): 417-458.

[210] Rockoff, Jonah. 2004. "The Impact of Individual Teachers on Student Achievement: Evidence from Panel Data." *The American Economic Review*, 94(2): 247-252.

[211] Rockoff, Jonah, Douglas Staiger, Thomas Kane, and Eric Taylor. 2012. "Information and Employee Evaluation: Evidence from a Randomized Intervention in Public Schools." *American Economic Review*, 102(7): 3184-3213.

[212] Rosenbaum, James. 1995. "Changing the Geography of Opportunity by Expanding Residential Choice: Lessons from the Gautreaux Program." *Housing Policy Debate*, 6(1): 231-269.

[213] Rothstein, Jesse and Till von Wachter. 2016. "Social Experiments in the Labor Market." *Handbook of Economic Experiments*, forthcoming.

[214] Rouse, Cecilia Elena. 1998. "Private School Vouchers and Student Achievement: An Evaluation of the Milwaukee Parental Choice Program." *The Quarterly Journal of Economics*, 113(2): 553-602.

[215] Ryan, Elizabeth McIntyre. 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.

[216] Ryan, Richard. 1982. "Control and Information in the Intrapersonal Sphere: An Extension of Cognitive Evaluation Theory." *Journal of Personality and Social Psychology*, 63: 397-427.

[217] Sanbonmatsu, Lisa, Jens Ludwig, Larry Katz, Lisa Gennetian, Greg Duncan, Ronald Kessler, Emma Adam, Thomas McDade, and Stacy Tessler Lindau. 2011. "Moving to Opportunity for Fair Housing Demonstration Program: Final Impacts Evaluation." Washington, D.C.: U.S. Department of Housing and Urban Development.

[218] Schmitt, John, and Jonathan Wadsworth. 2006. "Changing Patterns in the Relative Economic Performance of Immigrants to Great Britain and the U.S., 1980-2000." Cambridge, MA: CEPR.

[219] Schultz, T. Paul. 2000. "Final Report: The Impact of PROGRESA on School Enrollments." Washington, D.C.: International Food Policy Research Institute.

[220] Schwartz, Robert. 2005. "Literacy Learning of At-Risk First-Grade Students in the Reading Recovery Early Intervention." *Journal of Educational Psychology*, 97(2): 257-267.

[221] Schweinhart, Lawrence, H. Barnes, and D.P. Weikart. 1993. "Significant Benefits: The HighScope Perry Preschool Study Through Age 27." Ypsilanti, MI: HighScope Press.

[222] Schweinhart, Lawrence, Jeanne Montie, Zongping Xiang, William Barnett, Clive Belfield, and Milagros Nores. 2005. "Lifetime Effects: The High/Scope Perry Preschool Study Through Age 40." Ypsilanti, MI: HighScope Press.

[223] Skoufias, Emmanuel. 2005. "PROGRESA and Its Impacts on the Welfare of Rural Households in Mexico." Washington, D.C.: International Food Policy Research Institute.

[224] Slavin, Robert. 2010. "Can Financial Incentives Enhance Educational Outcomes? Evidence from International Experiments." *Educational Research Review*, 5(1): 68-80.

[225] Slavin, Robert, Marshall Leavey, and Nancy Madden. 1984. "Combining Cooperative Learning and Individualized Instruction: Effects on Student Mathematics Achievement, Attitudes, and Behaviors." *The Elementary School Journal*, 84(4): 409-422.

[226] Springer, Matthew, Dale Ballou, Laura Hamilton, Vi-Nhuan Le, J.R. Lockwood, Daniel F. McCaffrey, Matthew Pepper, and Brian M. Stecher. 2010. "Teacher Pay for Performance." Nashville, TN: NCPI.

[227] Springer, Matthew, John Pane, Vi-Nhuan Le, Daniel F. McCaffrey, Susan Burns, Laura Hamilton, and Brian M. Stecher. 2012. "Team Pay for Performance." *Educational Evaluation and Policy Analysis*, 34(4): 367-390.

[228] St. Pierre, Robert, Jean Layzer, Barbara Goodson, and Lawrence Bernstein. 1997. "National Impact Evaluation of the Comprehensive Child Development Program." Cambridge, MA: Abt Associates.

[229] Sumi, W., Michelle Woodbridge, Harold Javitz, Patrick Thornton, Mary Wagner, Kristen Rouspil, Jennifer Yu, John Seeley, Hill Walker, Annemieke Golly, Jason Small, Edward Feil, and Herbert Severson. "Assessing the Effectiveness of First Step to Success: Are Short-Term Results the First Step to Long-Term Behavioral Improvements?" *Journal of Emotional and Behavioral Disorders*, 21(1): 1-14.

[230] Taylor, Eric, and John Tyler. 2012. "The Effect of Evaluation on Teacher Performance." *The American Economic Review*, 102(7): 3628-3651.

[231] The New Teacher Project (TNTP). 2015. "The Mirage: Confronting the Hard Truth About Our Quest for Teacher Development." Brooklyn, NY: The New Teacher Project.

[232] Todd, Petra, and Kenneth Wolpin. 2003. "On the Specification and Estimation of the Production Function for Cognitive Achievement." *The Economic Journal*, 113(485): F3-F33.

[233] Tucker, Marc. 2011. "Teacher Quality: What's Wrong with U.S. Strategy?" *Educational Leadership*, 49(4): 42-46.

[234] Tuttle, Christina, Brian Gill, Phillip Gleason, Virginia Knechtel, Ira Nichols-Barrer, and Alexandra Resch. 2013. "KIPP Middle Schools: Impacts on Achievement and Other Outcomes. Final Report." Princeton, NJ: Mathematica Policy Research.

[235] U.S. Department of Education. 2009. "State and Local Implementation of the No Child Left Behind Act." Washington, D.C.: U.S. Department of Education.

[236] U.S. Department of Education. 2014. "Fiscal Year 2015 Education Budget Summary and Background Information." Washington, D.C.: U.S. Department of Education.

[237] U.S. Department of Education. 2015. Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, What Works Clearinghouse.

[238] U.S. Government Accountability Office. 2014. "K-12 Education: Characteristics of the Investing in Innovation Fund." Washington, D.C.: U.S. Government Accountability Office.

[239] United Nations Development Programme. "Human Development Report 2010." New York.: United Nations.

[240] Vigdor, Jacob, and Helen Ladd. 2010. "Scaling the Digital Divide: Home Computer Technology and Student Achievement." NBER Working Paper no. 16078.

[241] von Loeffelholz, Hans Dietrich. 2011. "Social and Labor Market Integration of Ethnic Minorities in Germany." In: Martin Kahanec, and Klaus Zimmerman, editors. *Ethnic Diversity in European Labor markets: Challenges and Solutions*. 109-136. Cheltenham, UK: Edward Elgar Publishing.

[242] Weikart, David, Dennis Deloria, Sarah Lawser, and Ronald Wiegerink. 1970. "Longitudinal Results of the Ypsilanti Perry Preschool Project." Ypsilanti, MI: High/Scope Educational Research Foundation.

[243] Wilson, Timothy, Patricia Linville. 1982. "Improving the Academic Performance of College Freshmen: Attribution Therapy Revisited." *Journal of Psychology and Social Psychology*. 42(2): 367-376.

[244] Winship, Scott, and Stephanie Owen. 2013. "The Brookings Social Genome Model." Washington, D.C.: Brookings.

[245] Witte, John. 1997. "Achievement Effects of the Milwaukee Voucher Program." Paper presented at the 1997 American Economics Association Annual Meeting, New Orleans, LA.

[246] Witte, John, Troy Sterr, and Christopher Thorn. 1995. "Fifth-Year Report Milwaukee Parental Choice Program." LaFollette School Working Paper no. 1995-001.

[247] Wolf, Patrick, Babette Gutmann, Michael Puma, Brian Kisida, Lou Rizzo, Nada Elissa, and Matthew Carr. 2010. "Evaluation of the DC Opportunity Scholarship Program." Washington, D.C.: National Center for Education Evaluation and Regional Assistance.

[248] Worrall, John. 2007. "Evidence in Medicine and Evidence-Based Medicine." *The Philosophy Compass*, 2(6): 981-1022.

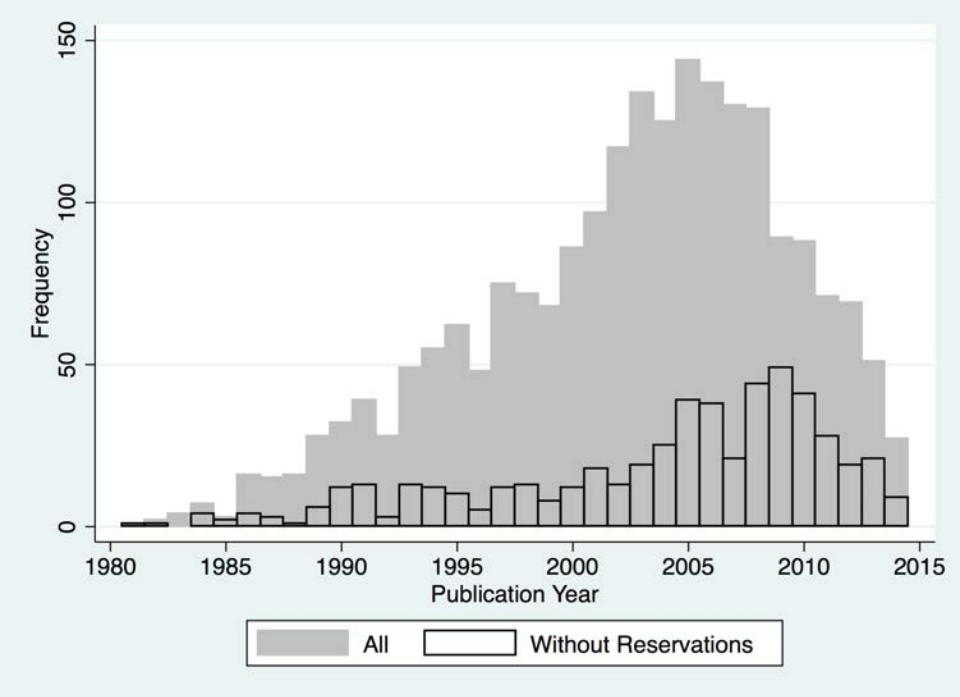
[249] Yeager, David, and Gregory Walton. 2011. "Social-Psychological Interventions in Education." *Review of Educational Research*, 81(2): 267-301.

[250] Yeates, Kieth, David MacPhee, Frances Campbell, and Craig Ramey. "Maternal IQ and Home Environment as Determinants of Early Childhood Intellectual Competence: A Developmental Analysis." *Developmental Psychology*, 19: 731-739.

[251] York, Benjamin, and Susanna Loeb. "One Step at a Time: The Effects of an Early Literacy Text Messaging Program for Parents and Preschoolers." NBER Working Paper no. 20659.

[252] Zvoch, Keith, and John Stevens. 2012. "Summer School Effects in a Randomized Field Trial." *Early Childhood Research Quarterly*, 28(1): 24-32.

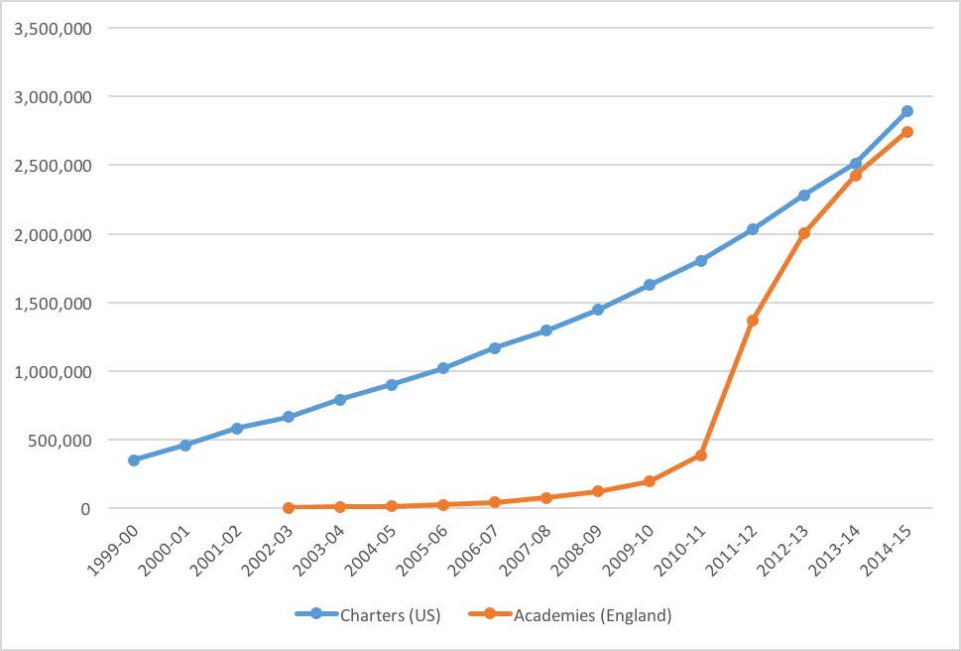
Figure 1
Reviewed WWC Studies



Notes: 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.

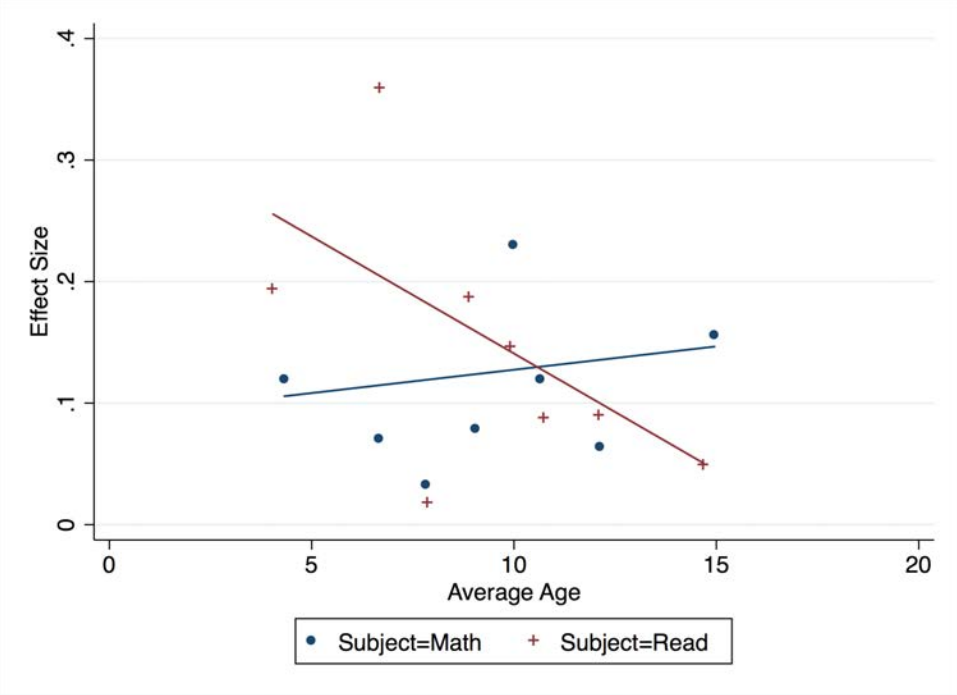
Figure 2

Number of Students in "Charters"



Notes: This figure presents the number of students enrolled in charter schools (U.S.) and academies (England) for the 1999-00 to 2014-15 school years. Academies are publicly funded independent schools. Similar to the U.S. charter schools, academies don't have to follow the national curriculum and term lengths. The U.S. data comes from the National Alliance for Public Charter Schools (NAPCS) and the data for England comes from the U.K. Department of Education. Note that the U.S. 2014-2015 number is estimated (NAPCS 2015).

Figure 3
Correlation of Effect Size and Average Age of Intervention



Notes: This figures 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.

Table 1: Paper Accounting

	Number of Papers
Panel A: Titles Found	(1)
From Broad Search	≈ 8,000
Selected for Further Review	859
TOTAL INCLUDED	196
Panel B: Reason For Exclusion	
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.

Table 2: Meta-Analysis

	Math			Reading		
	Unweighted Average	Fixed Effects	Random Effects	Unweighted Average	Fixed Effects	Random Effects
Panel A: Early Childhood	(1)	(2)	(3)	(4)	(5)	(6)
ALL	0.120	0.111	0.111	0.202	0.106	0.189
	(0.028)	(0.031)	(0.031)	(0.027)	(0.012)	(0.027)
		20			44	
Panel B: Home						
ALL	0.039	-0.004	-0.004	0.078	0.010	0.010
	(0.045)	(0.008)	(0.008)	(0.052)	(0.007)	(0.007)
		8			22	
Parental Involvement	0.122	-0.001	-0.001	0.143	0.009	0.034
	(0.115)	(0.021)	(0.021)	(0.103)	(0.021)	(0.050)
		3			11	
Educational Resources	-0.060	-0.060	-0.060	0.072	0.015	0.015
	(0.000)	(0.050)	(0.050)	(0.063)	(0.014)	(0.014)
		1			7	
Poverty Reduction	0.008	0.008	0.008	0.022	0.016	0.016
	(0.001)	(0.029)	(0.029)	(0.011)	(0.024)	(0.024)
		2			4	
Panel C: School						
ALL	0.135	0.035	0.053	0.203	0.023	0.069
	(0.022)	(0.004)	(0.009)	(0.028)	(0.004)	(0.011)
		72			98	
Student Incentives	0.039	0.016	0.024	0.097	0.016	0.021
	(0.026)	(0.011)	(0.018)	(0.072)	(0.011)	(0.017)
		5			8	
High Dosage Tutoring	0.393	0.309	0.309	0.405	0.217	0.229
	(0.095)	(0.106)	(0.106)	(0.047)	(0.030)	(0.033)
		6			25	
Low Dosage Tutoring	0.074	0.015	0.015	0.050	0.015	0.015
	(0.045)	(0.013)	(0.013)	(0.045)	(0.015)	(0.015)
		3			4	
Teacher Certification	0.031	0.028	0.030	0.000	0.007	0.007
	(0.036)	(0.012)	(0.030)	(0.015)	(0.028)	(0.028)
		5			3	
Teacher Incentives	0.052	0.002	0.022	-0.000	-0.006	-0.006
	(0.033)	(0.011)	(0.022)	(0.021)	(0.012)	(0.012)
		7			4	
General PD	0.173	0.019	0.019	0.153	0.022	0.022
	(0.075)	(0.024)	(0.024)	(0.060)	(0.023)	(0.023)
		7			9	
Managed PD	0.059	0.052	0.052	0.493	0.217	0.403
	(0.009)	(0.016)	(0.016)	(0.187)	(0.029)	(0.120)

		2			8	
Data-Driven	0.107	0.043	0.057	0.071	0.009	0.030
	(0.041)	(0.014)	(0.024)	(0.040)	(0.011)	(0.024)
		4			4	
Extended Time	-0.033	0.019	-0.019	0.155	0.012	0.032
	(0.089)	(0.026)	(0.068)	(0.136)	(0.029)	(0.048)
		4			5	
School Choice/Vouchers	0.076	0.024	0.024	0.070	-0.010	0.023
	(0.035)	(0.018)	(0.018)	(0.040)	(0.012)	(0.025)
		6			7	
Charters	0.121	0.088	0.110	0.072	0.038	0.048
	(0.039)	(0.011)	(0.030)	(0.026)	(0.010)	(0.018)
		9			9	
No Excuse Charters	0.170	0.124	0.153	0.104	0.055	0.077
	(0.048)	(0.022)	(0.042)	(0.040)	(0.018)	(0.031)
		5			5	

Notes: This table reports average effects for categories of papers discussed in the main text. Columns (1)-(3) report reults 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.