

---

## Running Randomized Evaluations

---

# Running Randomized Evaluations

A PRACTICAL GUIDE

RACHEL  
GLENNERSTER    *and*    KUDZAI  
TAKAVARASHA

---

PRINCETON UNIVERSITY PRESS

*Princeton and Oxford*

Copyright © 2013 by Princeton University Press

Published by Princeton University Press, 41 William Street, Princeton, New Jersey 08540  
In the United Kingdom: Princeton University Press, 6 Oxford Street, Woodstock, Oxfordshire  
OX20 1TW

press.princeton.edu

Cover design by Leah E. Horgan

All Rights Reserved

Library of Congress Cataloging-in-Publication Data

Glennerster, Rachel.

Running randomized evaluations : a practical guide / Rachel Glennerster  
and Kudzai Takavarasha.

pages      cm

Includes bibliographical references and index.

ISBN 978-0-691-15924-9 (hardcover : alk. paper) — ISBN 978-0-691-15927-0  
(pbk. : alk. paper)

1. Evaluation research (Social action programs) 2. Social sciences—Research.

I. Takavarasha, Kudzai, 1973– II. Title.

H61.G5544 2013

001.4'34—dc23

2013014882

British Library Cataloging-in-Publication Data is available

This book has been composed in Minion Pro with ITC Franklin Gothic Display by  
Princeton Editorial Associates Inc., Scottsdale, Arizona

Printed on acid-free paper. ∞

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

## CONTENTS

*Preface* vii

*Abbreviations and Acronyms* ix

- 1 The Experimental Approach 1
- 2 Why Randomize? 24
- 3 Asking the Right Questions 66
- 4 Randomizing 98
- 5 Outcomes and Instruments 180
- 6 Statistical Power 241
- 7 Threats 298
- 8 Analysis 324
- 9 Drawing Policy Lessons 386

*Appendix* 421

*Glossary* 443

*Index* 453

---

## 4 Randomizing



There are many different ways to introduce randomization into a program. In this chapter we discuss how to decide which approach is best in a given context. We also cover the process and mechanics of randomization. This chapter has six modules:

**MODULE 4.1: Opportunities to Randomize**

**MODULE 4.2: Choosing the Level of Randomization**

**MODULE 4.3: Deciding Which Aspects of the Program to Randomize**

**MODULE 4.4: The Mechanics of Simple Randomization**

**MODULE 4.5: Stratified and Pairwise Randomization**

**MODULE 4.6: A Catalog of Designs**

---

### **MODULE 4.1 Opportunities to Randomize**

*In this module we explain the three aspects of programs (access, timing, and encouragement to take up the program) that can be randomly assigned to create treatment and comparison groups. We then outline nine common situations that lend themselves to randomized evaluation, illustrated with examples from recent evaluations.*

#### **What can be randomized?**

In order to evaluate the impact of a program or policy, our randomly selected treatment group must have more exposure to the program than the comparison group. We can control three aspects of the program or policy to create this differential exposure:

1. Access: We can choose which people will be offered access to the program.
2. Timing of access: We can choose when to provide access to the program.
3. Encouragement: We can choose which people will be given encouragement to participate in the program.







Because we control these three aspects of the program, these are also the three aspects we can randomly allocate. Whether we vary access, timing of access, or encouragement to take part, we can vary each aspect by individual or by group. For example, we can randomize individuals within a community, offering access to some people but not to others, or we can randomize communities, offering access to all individuals in the chosen community.

Randomly assign the offer to access the treatment

Of the three possibilities discussed in this module, randomly assigning access to a program or policy is the most common. Imagine that we have enough resources to provide textbooks to only 100 schools. We would make a list of 200 eligible schools and then randomly select 100 to receive the textbooks during the evaluation period and then deliver the books to only these schools. The remaining 100 schools form the comparison group.

Randomly assign a time when people can access the treatment

We can randomly assign the time when people can access the program, designating who gets access first and who gets it later. Imagine a school-based deworming program in Kenya is planning to phase in their program to schools over three years. There are 75 eligible schools. We can randomly divide them into three groups of 25 each and randomly select which group starts the program in each of the three years (Figure 4.1). In the first year, Group A starts the program, and together Groups B and C form the comparison group. In year 2, Group B starts the program and, together with Group A, will make up the treatment group, while Group C is the comparison group. In year 3, Group C starts the program, and with all three groups now receiving treatment, there is no longer a comparison group. When we vary the timing of access, the difference in exposure to the program is created by delaying access for some people, as in this deworming exam-

Year	Group A		Group B		Group C
Year 1		Treatment group	Comparison group		Comparison group
Year 2		Treatment group		Treatment group	Comparison group
Year 3					

Evaluation ends  
No comparison group exists anymore

**FIGURE 4.1** Random assignment of the timing of the treatment through phase-in design

Note: The pill bottles indicate which groups had received the treatment in each of the three years.

ple. It can also be created, as we will see in Module 4.5, by having people take turns in receiving access.

Randomly assign encouragement to take up the program

Sometimes we are unable to randomly assign access to a program itself, but we can randomly assign encouragement to take up the program. This approach is useful when we want to evaluate a program that is already open to all of the eligible recipients but only some are currently using it. The program continues to be open to all the eligible people, but only some of them (the treatment group) will receive extra encouragement to take up the program.

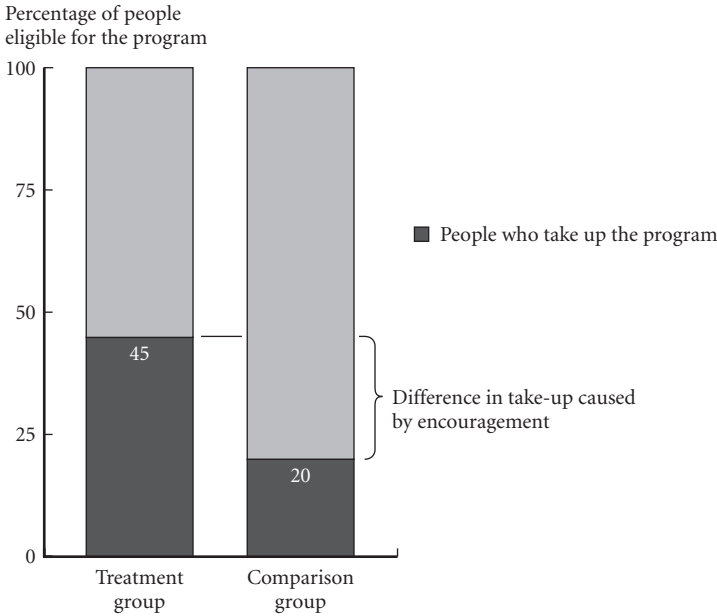
Imagine that we are evaluating a program that provides savings accounts for farmers growing cash crops. Any household that grows cash crops can sign up for an account, but only 50 of the 250 eligible households in the community have signed up for a savings account. It turns out that these 50 are much more likely to invest in inputs, and we want to know whether this is because the savings accounts help them save for their inputs. There are 200 eligible households remaining. We can split these into a treatment and a comparison group and give extra encouragement to open a savings account to the treatment group households. We send letters to these 100 randomly selected households telling them about the benefits of a savings account and offering to help them fill out the paperwork to open an account. For the other 100 households we do nothing. The hope is that a higher proportion of the encouraged households (the treatment group) will open an account. If, for example, 45 of the 100 households in the treatment group take up the program and 20 in the status quo comparison

group take it up, the treatment group will have a higher proportion enrolled in the program—45 percent compared to 20 percent—and this gives us the difference in exposure to the program needed to measure program impact (Figure 4.2).

We can think of encouragement as making it easier for some people to take up a program. Because it is easier for them, they are more likely to take up the program, and we have the difference in exposure we need.

***When is it possible to perform a randomized evaluation?***

In general, opportunities to randomize arise when implementers want to design a new approach to addressing a problem, when a new program is being introduced, and when there are insufficient resources to provide a service to all those who could benefit. We cover 10 of the most common examples in Table 4.1. In most cases, conducting a randomized evaluation will not change the number of people who receive the program, and this is the case for all the opportunities listed here.



**FIGURE 4.2** Encouragement as motivation for more people in the treatment group to take up the program



**TABLE 4.1** Opportunities to randomize

Opportunity	Description
New program design	When a problem has been identified but there is no agreement about what solution to implement. The evaluators work with the implementers from the outset to design programs and then pilot-test them.
New programs	When a program is new and being pilot-tested.
New services	When an existing program offers a new service.
New people	When a program is being expanded to a new group of people.
New locations	When a program is being expanded to new areas.
Oversubscription	When there are more interested people than the program can serve.
Undersubscription	When not everyone who is eligible for the program takes it up.
Rotation	When the program's benefits or burdens are to be shared by rotation.
Admission cutoffs	When the program has a merit cutoff and those just below the cutoff can be randomly admitted.
Admission in phases	When logistical and resource constraints mean that not all the potential beneficiaries can be enrolled at one time and people can be randomly admitted in phases over time.

(See Module 2.4 for a discussion of ethics on this point.) There are cases on this list in which doing a randomized evaluation changes the approach to targeting under the program, and we discuss the ethical implications of this in Module 4.3.<sup>1</sup>

#### New program design

We may have identified a problem and concluded that it requires a new solution. In such cases, the evaluators can start working with the

1. For a more detailed discussion of ethics in randomized evaluations, see Glennerster and Powers in *The Oxford University Press Handbook on Professional Economic Ethics* (Oxford, UK: Oxford University Press, forthcoming), and William R. Shadish, Thomas D. Cook, and Donald T. Campbell, *Experimental and Quasi-experimental Designs for Generalized Causal Inference* (Boston: Houghton Mifflin, 2002).

implementing organization before there is a program or a set of alternatives to test and contribute to their design.

*Example: A research partnership pilots a series of health programs.* The NGO Seva Mandir has been working for almost 60 years in the Udaipur District of Rajasthan, India. Its programs target education, the environment, microfinance, and health. In 2002 Seva Mandir partnered with researchers on a comprehensive survey to identify major health and healthcare needs. They jointly developed and pilot-tested three interventions in a series of randomized evaluations.<sup>2</sup>

### New programs

When a program is new and its effects are unknown, a randomized evaluation often is the best way to estimate its impact. Randomly selecting the beneficiaries of a pilot program is also often perceived as the fairest way to allocate the program.

*Example: The government pilots a new social welfare program in Mexico.* The nationwide CCT program in Mexico was first introduced as a pilot program in rural areas of seven states under the name PROGRESA. Of the 506 communities that were sampled, 320 were randomly assigned to receive the pilot program. Those assigned to the comparison group received the program only after the evaluation, when the program had been found to be effective and was scaled up.<sup>3</sup>

### New services

Programs evolve to meet new challenges. They can, for example, change the pricing or method of delivery, or they can add altogether new services. Much as in the case of the pilot test, the innovation can be randomly assigned to existing or new clients for evaluation.

*Example: Microfinance institutions add new services.* Besides credit for the poor, many microfinance institutions (MFIs) now offer additional services such as savings accounts, health insurance, and business training. For example, an MFI in India introduced health insurance; in the Philippines, an MFI introduced a loan product with individual liability in addition to its group liability product; and in India, an MFI introduced financial literacy training.<sup>4</sup> The rollout of

2. See Module 3.3 for further details.

3. This study by T. Paul Schultz is summarized as Evaluation 9 in the appendix.

4. Many of these studies are covered by Jonathan Bauchet, Cristobal Marshall, Laura Starita, Jeanette Thomas, and Anna Yalouris in “Latest Findings from Randomized Evaluations of Microfinance,” in *Access to Finance Forum* (Washington, DC: Con-

these new services to existing customers was randomized, creating an opportunity for evaluation.

*Example: An HIV education program in Kenya adds a risk information campaign.* International Child Support (ICS, a Netherlands-based NGO) in Kenya piloted a program to improve the delivery of the national HIV education curriculum. The program had three existing components. Later an information campaign was added to address the danger of cross-generational sexual activity. This component was randomly assigned to 71 of the 328 participating schools.<sup>5</sup>

#### New people and locations

Programs expand by adding new people in their existing locations or moving to new locations. When there are not enough resources to cover all the new clients at once, randomizing may be the fairest way to decide who will be served first.

*Example: A US state extends Medicaid to those just above the usual Medicaid cutoff.* US state governments, with federal support, provide health insurance to millions of poor families. The state of Oregon had limited funding to extend the pool of people who were eligible for this program. They held a lottery among eligible low-income individuals to decide who would be offered services. Later researchers compared the outcomes of lottery winners and losers in terms of healthcare use, health status, and financial strain.<sup>6</sup>

*Example: A remedial education program is expanded to a new city.* In 2000 the Indian organization Pratham expanded their remedial education program to the city of Vadodara. There were 123 schools in the city, and each school received tutors, but the tutor was randomly assigned to either grade 3 or grade 4.<sup>7</sup>

---

sultative Group to Assist the Poor, 2011), <http://www.cgap.org/gm/document-1.9.55766/FORUM2.pdf>. For the smoking study in the Philippines, see Dean Karlan, Xavier Gine, and Jonathan Zinman, "Put Your Money Where Your Butt Is: A Commitment Savings Account for Smoking Cessation," *American Economic Journal: Applied Economics* 2 (2010): 213–235.

5. This study by Esther Duflo, Pascaline Dupas, and Michael Kremer is summarized as Evaluation 4 in the appendix.

6. Amy Finklestein, Sarah Taubman, Bill Wright, Mira Bernstein, Jonathan Gruber, Joseph P. Newhouse, Heidi Allen, Katherine Baicker, and Oregon Health Study Group. 2012. "The Oregon Health Insurance Experiment: Evidence from the First Year." *Quarterly Journal of Economics* 127 (3): 1057–1106.

7. This study by Abhijit Banerjee, Shawn Cole, Esther Duflo, and Leigh Linden is summarized as Evaluation 2 in the appendix.

### Oversubscription

Often in developing countries, more people are interested in a program than the program has resources to serve. When demand outstrips supply, random assignment may be the fairest way to choose the participants.

*Example: The government uses a lottery to decide who receives secondary school tuition vouchers in Colombia.* A program in Colombia provided vouchers for private school to students from poor families. There were more qualified and interested students than vouchers. The municipalities assigned the vouchers by lottery to ensure fairness.<sup>8</sup>

*Example: The US government uses a lottery to decide who receives vouchers for housing in more affluent neighborhoods.* A pilot program in the United States provided vouchers for families to move from high-poverty housing projects to low-poverty neighborhoods. Demand for housing assistance in the United States often exceeds the supply of public housing available. The US Department of Housing and Urban Development randomly offered vouchers to families in order to rigorously evaluate the program and distribute the limited vouchers in a way that was fair.

### Undersubscription

Sometimes program take-up is low and a program is serving fewer people than it could, even though there are enough resources to cover everyone and the program is open to everyone. One way to increase demand is to offer more encouragement to take up the program through either more information or more incentives. This additional encouragement can be randomized. Necessary but unpopular programs may also be undersubscribed (as in the case of the military draft for the Vietnam War or the quotas for women in politics in India). Lotteries are often seen as a fair way to distribute the burden of unpopular programs.

*Example: A large American university encourages employees to sign up for retirement savings.* Like many employers in the United States, a large American university has a retirement savings program. Every year the university organizes benefits fairs to provide information and enroll employees. One year the university offered a reward of

8. This study by Eric Bettinger, Michael Kremer, and Juan E. Saavedra is summarized as Evaluation 10 in the appendix.

\$20 for attending the fair. This extra encouragement was randomly allocated.<sup>9</sup>

*Example: A large South African lender encourages borrowers through direct mail.* The lender provides loans to high-risk borrowers at a high interest rate. The firm used direct mailings to advertise its products and increase demand. The recipients and the content of the letters were randomized to measure the effects on demand.<sup>10</sup>

*Example: The US government had too few volunteer soldiers during the Vietnam War.* Young men who were eligible for the draft in the United States were entered into a lottery to determine whether they would have to serve in the armed services. Subsequently, academics used the lottery to determine the impact of serving in the armed services on future labor market outcomes.<sup>11</sup>

### Rotation

Sometimes there are just enough resources to cover a portion of the people potentially eligible for a program, and the resources will remain fixed for a long time. Because everyone wants access to the program and they cannot all have it at once, one approach is to have people take turns. The order in which program resources or burdens rotate can be decided randomly, with the group in rotation serving as the treatment group. With this approach we have to worry about effects lingering after the program has rotated away to another group, improving or worsening outcomes in the next period.

*Example: Some state governments in India randomly cycle political quotas among village councils.* In India, a third of all rural village councils in each election cycle must elect a woman as president. To ensure that the reservations rotate fairly among the councils, some states determine the reservation schedule randomly.<sup>12</sup>

9. This study by Esther Duflo and Emmanuel Saez is summarized as Evaluation 11 in the appendix.

10. Dean Karlan and Jonathan Zinman, "Observing Unobservables: Identifying Information Asymmetries with a Consumer Credit Field Experiment," *Econometrica* 77 (2009): 1993–2008.

11. Joshua D. Angrist, "Lifetime Earnings and the Vietnam Era Draft Lottery: Evidence from Social Security Administrative Records," *American Economic Review* 80 (1990): 313–336.

12. This study by Lori Beaman, Raghavendra Chattopadhyay, Esther Duflo, Rohini Pande, and Petia Topalova is summarized as Evaluation 12 in the appendix.

### Admission cutoffs

Some programs have admission cutoffs. These cutoffs may matter for program targeting, but the precise cutoff point is often somewhat arbitrary. For example, women with less than 1 acre of land may be eligible for special assistance. But how much less deserving of government support is a woman with 1.05 acres of land compared to one with 0.95 acre? Somewhat arbitrary cutoffs give us an opportunity to randomize.

*Example: A rural bank in the Philippines uses credit scores.* Eligibility for business loans at First Macro Bank in the Philippines depended on a credit score ranging from zero to 100. Those with scores of 30 and below were considered uncreditworthy and automatically rejected, and those with scores of 60 and above were considered very creditworthy and approved. A random sample of applications with scores between 31 and 59 (i.e., who were marginally creditworthy) were approved and formed the treatment group for the evaluation, while the rest of those in this range were rejected and formed the comparison group. The evaluation assessed the impact of giving credit to those who were marginally creditworthy.<sup>13</sup>

### Admission in phases

Sometimes a program does not have the resources or logistical capacity to deliver the program to everyone at once. If resources are going to grow, then people may be admitted into the program as the resources become available. This presents an opportunity to randomize the order in which people are phased in.

*Example: An NGO introduced deworming in phases.* Because of logistical and financial constraints, ICS Africa could add only 25 schools a year to their deworming program in Kenya. The schools were randomly divided into three groups of 25, and one (randomly selected) group was phased in each year. At the end of three years, all the schools had been treated.<sup>14</sup>

13. This study by Dean Karlan and Jonathan Zinman is summarized as Evaluation 13 in the appendix.

14. This study by Sarah Baird, Joan Hamory Hicks, Michael Kremer, and Edward Miguel is summarized as Evaluation 1 in the appendix.

## Module 4.1 summary

*What aspects of a program can be randomized?*

- Three aspects of a program can be randomized:
  1. Access to the program
  2. Timing of access to the program
  3. Encouragement to take up the program
- Randomizing access to the program is the most common of the three possibilities.
- We randomize the timing of access to a program when the program cannot reach all the people at the same time. The people are then admitted in phases. If the program effects are short-term and transient, the program can rotate access among the eligible.
- When we randomize encouragement to take up a program, we offer additional information or incentives to some of the people. We do this mostly when the program is open to all eligible people but is undersubscribed.

*When do opportunities to randomize commonly arise?*

- When the program is being pilot-tested
- When the program is adding new services
- When the program is adding new people
- When the program is adding new locations
- When the program is oversubscribed
- When the program is undersubscribed
- When the program has to be shared through rotation
- When the program has an admission cutoff
- When the program has to admit people in phases

---

## MODULE 4.2 Choosing the Level of Randomization

*This module describes how we decide whether to randomize individuals, households, communities, schools, clinics, or other units.*

Most programs and policies seek to change the lives of individuals, but they often do so by working through groups such as schools, savings groups, households, or communities. When we design our randomized evaluation we have to decide whether to randomize individuals in or out of a program or to randomize whole groups in or out of the program.

Figure 4.3 describes the different steps involved in an individual-level and a group-level randomization. Panel A shows the steps involved in selecting a random sample of women farmers for involvement in an agricultural program in Sierra Leone. In partnership with those implementing the program we select from the three districts in Sierra Leone in which the program has the potential to operate three villages in which to run the program and the evaluation. Within these communities we determine who is eligible for the program (in this example, adult women). The women in these communities are then randomized into treatment and comparison groups. Those in the treatment group are offered access to the program, and those in the comparison group are not. Data are collected on all the women whether they are allocated to the treatment or the comparison group.

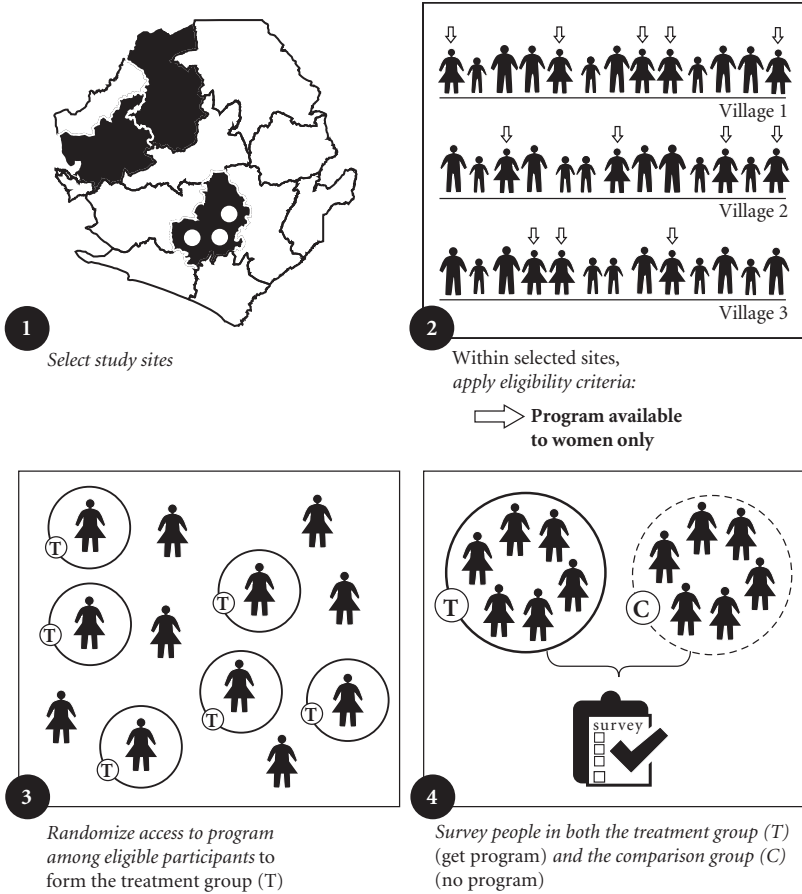
For a group-level randomization we again start by selecting a study site. As we discuss later in this chapter, we may want to choose our study sites randomly to ensure that the study is representative of a large geographic area. In this example we randomly pick three districts to be representative of the country as a whole. Again we use eligibility criteria to select participation in the program. We select medium-size villages (smaller villages will have few farmers who can benefit from our training, and large villages tend to have a higher ratio of non-farmers). Among these communities we randomly select some to receive the training and others not to. In this case we do not need to survey everyone in the community; we survey only a random sample of eligible people within the community.

The most common levels at which to randomize are the individual and the community level, but we can also randomize by school, health clinic, agricultural cooperative, or grade.

Some programs operate at several levels. Many microcredit organizations make loans to individuals through lending groups that are part of centers. These centers are visited by credit officers who are responsible for several centers. For an evaluation we could randomize by individual, lending group, village, or credit officer. The higher the level



A Individual-level randomization



**FIGURE 4.3** Randomization steps for individual- versus group-level randomization

of randomization, the larger the number of people who are randomized together as a group.

Usually the unit of randomization is determined by the unit at which the program is implemented. A program that works through clinics will usually be randomized at the clinic level—with some clinics receiving additional support and others not. This is a matter of practicality: it is usually infeasible to prevent some individuals from having access to a program that is implemented at the community

B Group-level randomization

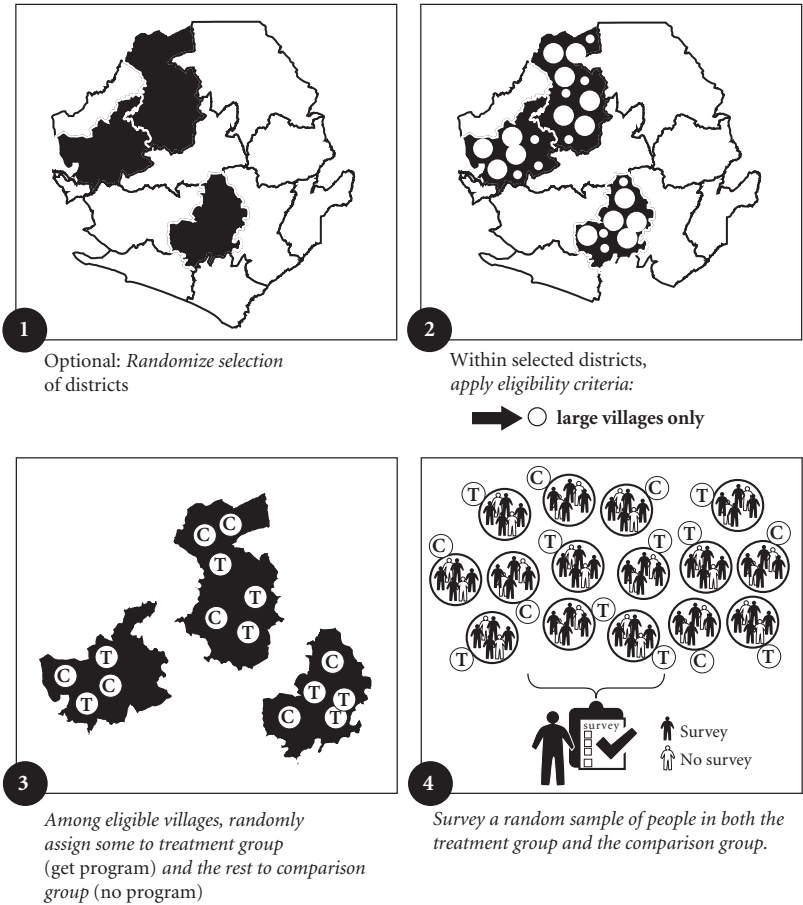


FIGURE 4.3 (continued)

level. But we do not always randomize at the level of implementation. As the microfinance example shows, a program may have many different levels at which it is implemented. But there are many other considerations. Six key considerations in determining the level of randomization are listed in Table 4.2.

**Unit of measurement and unit of randomization**

The level of randomization needs to be the same or higher than the level at which our outcome measure will be taken. Imagine that we are

**TABLE 4.2** Technical and nontechnical considerations for choosing the level of randomization

Consideration	Question to be answered
Unit of measurement	What is the unit at which our outcome will be measured?
Spillovers	Are there spillovers? Do we want the impact estimate to capture them? Do we want to measure them?
Attrition	Which level is best to keep participants from dropping out of the sample?
Compliance	Which level is best to keep participants from dropping out of the sample?
Statistical power	Which level gives us the greatest probability of detecting a treatment? If we randomize at the group level, will we still have a large enough effective sample size?
Feasibility	Which level is feasible ethically, financially, politically and logistically? Which is easiest, and which is least costly?

interested in the effects of worker retraining on firm profits. We could randomize training at the worker level, but we can measure the outcome of interest (profits) only at the firm level. We therefore need to randomize at the firm level.

Similarly, a voter education campaign in India (implemented with the help of two local NGOs, the Sarathi Development Foundation and Satark Nagrik Sangathan) provided information on the quality of candidates. The evaluation authors were interested in the impact of this information on electoral outcomes. Voting in India is anonymous, so it is possible to collect only self-reported data on individual voting. The authors were concerned that self-reported voting data might be systematically different from actual voting—for example, that certain types of people might be unwilling to report how they voted. Actual voting reports are available only at the aggregate polling station level. So even though the treatment targeted individual households, the campaign was randomized at the polling station level.<sup>15</sup>

15. Abhijit Banerjee, Donald Green, Jennifer Green, and Rohini Pande, “Can Voters Be Primed to Choose Better Legislators? Experimental Evidence from Rural India,” working paper, Massachusetts Institute of Technology, Cambridge, MA, 2009.

### *Spillovers and unit of randomization*

Programs can have direct and indirect effects. The indirect effects are called *spillovers* or *externalities*. Spillovers can take many forms and can be positive or negative. They can occur through a number of channels:

- *Physical*: Children immunized by a program reduce disease transmission in their community. Farmers who learn pig husbandry from a program increase pollution.
- *Behavioral*: A farmer imitates the fertilizer application techniques her neighbor in a treatment group learned from a program.
- *Informational*: People learn about the effectiveness of insecticide-treated bed nets from others who received these nets through a program (also known as social learning).
- *Marketwide or general equilibrium*: Older workers lose their jobs because firms receive financial incentives from a program to hire young workers.

#### Why spillovers matter

The formal requirement for a randomized evaluation to be valid is that the outcome of one person does not depend on the group to which other people around that person are allocated. This assumption does not hold when we have spillovers. A farmer may be more likely to use fertilizer if his neighbor is allocated to the treatment group than to the comparison group because the farmer will observe his neighbor using fertilizer and copy her example.

The problem is that when an individual's actions depend on whether someone else was assigned to the treatment or the comparison group, the difference in outcomes between the treatment and the comparison group no longer represents the impact of the program. For example, if comparison group farmers learn about the benefits of fertilizer from their treatment group neighbors, these comparison group farmers no longer represent a good counterfactual: they no longer behave in a way that the treatment group farmers would have behaved if the program had not existed.

Choosing the level of randomization to limit spillovers to the comparison group

If the treatment and comparison groups do not interact, there will be no spillovers. For spillovers to occur, the different groups must have

something in common, which can act as a transmission channel. All things equal, the greater the physical proximity and interaction between the treatment and comparison groups, the greater the spillovers. Choosing a unit of randomization so that most relevant interactions occur between people in the same group is the best way to limit spillovers to the comparison group.

Take the example of the relative risk HIV education program in Kenya. During a 40-minute class period, a program officer from ICS Africa shared and discussed information on the relative risk of HIV by age and gender. There was concern that children who attended the class would talk to their friends in other classes about what they had learned. To reduce the risk of spillovers, randomization took place at the level of the school.<sup>16</sup> The farther apart the schools were from each other, the more likely that the program effects would all be contained within the school where the program was offered.

Choosing the level at which to measure isolated spillover effects

If we want to measure spillover effects, we need within the comparison group a “spillover” group and a “no spillovers” group. In our analysis we compare outcomes for those in the treatment group with those in the no spillovers group to assess the impact of the program. By comparing outcomes from the spillover group and the no spillovers group we will assess the level of spillovers (Module 8.2). In choosing the level of randomization, we need to allow for this variation in exposure to spillovers. If we choose too high a level of randomization, there will be no contact between the treatment group and the comparison group. We will have eliminated the problem of spillovers but will not be able to measure them. If we choose too low a level of randomization, all of our comparison group will experience spillovers and we will not be able to evaluate the program because we will have no “pure” comparison.

We generate variation in exposure to spillovers by recognizing that the extent of spillovers will depend on the proportion of those around an individual who are treated. Consider physical spillovers. A cold is going around your office. The likelihood that you will catch that cold and be forced to take a sick day will depend on the number of co-workers with the cold. The same is true in the case of information

16. This study by Esther Duflo, Pascaline Dupas, and Michael Kremer is summarized as Evaluation 4 in the appendix.

spillovers: the higher the proportion of people in your circle who know a secret, the more likely you are to hear about it. This proportion of “treated” people is called the *treatment density*. We cover how to vary treatment densities in the last module of the chapter, but for now let’s look only at what matters in choosing the level of randomization.

*Untreated individuals in a treated group.* In the HIV information campaign discussed above, only children in the final grade of primary school (grade 8) were treated, but it is possible that they shared what they learned with other adolescents in the grades below (grades 6 and 7). Comparing the outcomes of grades 6 and 7 girls in treatment schools with the outcomes of grades 6 and 7 girls in comparison schools enables us to measure the spillover effects: although both sets of girls were untreated, the first set was exposed to the program through spillovers.

*Untreated units near treated units.* In most cases, the nearer untreated units are to treated units, the more likely they are to be affected by spillovers. The evaluation of school-based deworming took this into account. Randomization was at the level of the school, but there was enough out-of-school mixing of children who went to different schools that a deworming program in one school had the potential to impact children in neighboring schools. The random variation generated by the phase-in design meant that some not-yet-treated schools were near treated schools, while others were not.

*Using two levels of randomization to measure spillovers.* If measuring spillovers is one of the main objectives of our evaluation, it may be useful to randomize at two different levels, usually the individual level and the group level. Take the HIV education example again. An alternative design would have been to randomly choose treatment schools where the HIV education would be given. Then one could have randomly picked some adolescents from these schools and called them to a meeting where the information on relative risk was given. The spillover group would then have been the adolescents in the school that did not attend the meeting, while the no spillovers group would have been adolescents at comparison schools who received no HIV education program.

### ***Attrition and unit of randomization***

*Attrition* is the term used when outcome data are missing from some of the people in the sample. This can happen because people drop out

of the study, refuse to answer some questions, or cannot be found by enumerators during the endline survey. The level of randomization can play a role in reducing attrition.

Randomizing within a group can reduce cooperation when people are not assigned their preferred treatment. No matter how fair the randomization procedure itself may be, it is sometimes difficult to avoid the impression that the program is picking and choosing unfairly. This can lead to resentment. People in the comparison group, for example, might be less likely to cooperate with data collection when they see others receiving benefits from participating while they receive nothing. Randomizing at a higher level has the potential to reduce resentment because people living side by side are treated similarly, and it might therefore help reduce attrition. In practice, however, most attrition results from people's moving or finding the survey too long and onerous to complete.

### ***Compliance and unit of randomization***

We want as few departures as possible from the study plan or protocol. We want the program staff to execute the experiment as planned, and we want the participants to adhere as closely as possible to their assigned treatment. Choosing the right level of randomization can increase compliance by both the program staff and the participants.

Increasing the likelihood of compliance by program staff

An evaluation design that is seen as unfair by program staff or that makes the job of the program staff too complicated is likely to run into problems. If a staff member is faced with two equally needy children and is told to provide the program to only one of them, he may end up sharing the resources between the children, undermining the validity of the experiment. Or if a credit officer has to provide one type of loan to some clients and another type to other clients, she may become confused and give the wrong type of loan to some clients. The solution is to make sure that program staff are not faced with these difficult choices or confusing situations by adjusting the level of randomization. For example, we may want to randomize at the program staff level so that one credit officer delivers only one version of the program.

Increasing compliance by participants

Study participants themselves can sometimes “undo” the randomization. Good design can minimize this risk. For example, the Work and

Iron Status Evaluation studied the impact of iron supplementation.<sup>17</sup> To select the participants, all the screened subjects were placed in a pool, and a sample of individuals was drawn. Once an individual was chosen, his entire household was included in the study sample. The decision to assign treatment or comparison status at the household level was motivated by two concerns. First, household members might share iron pills with other members of the household. At a more pragmatic level, many of the older respondents in the study had limited literacy and there was concern that it would be difficult for respondents to keep track of which pills were to be taken by whom if not all household members were to be taking them.

In Module 7.1 we discuss other strategies to minimize noncompliance as well as the importance of measuring the level of noncompliance. In Module 8.2 we discuss analysis of results if there is noncompliance.

### ***Statistical power and level of randomization***

All things equal, the larger the number of units that are randomized, the higher the statistical power. When we choose to randomize groups such as schools or villages rather than individuals, we have fewer units to randomize. Even though the program may still serve the same number of individuals and we may collect the same amount of data as if the unit of randomization was the individual, the statistical power is likely to be dramatically reduced. This is because the outcomes of people in the same unit (the same village or same school) are not fully independent of each other. We will discuss this in much more detail in Chapter 6, but a rough rule of thumb is that for statistical power, the number of units randomized is more important than how many people are interviewed.

### ***Feasibility and unit of randomization***

We may have good technical reasons to randomize at one level over another, but in the end we have to ask if it is feasible. There are at least four factors to consider:

1. Ethics: Is the randomization ethical?

17. Duncan Thomas et al., "Causal Effect of Health on Labor Market Outcomes: Experimental Evidence," online working paper series, California Center for Population Research, University of California–Los Angeles, 2006, <http://www.escholarship.org/uc/item/0g28k77w>.



2. Politics: Is it permitted? Does the community consent? Is it perceived as fair?
3. Logistics: Can we carry out the tasks of delivering this program at this level?
4. Cost: Do we have the money? Is this the best use of the money?

#### Ethics and feasibility

Module 2.4 sets out a commonly accepted framework for assessing the ethics of a given research design. We must respect the people involved in the trial and their wishes, balance any risks against likely benefits of doing the evaluation, and ensure that the benefits go at least in part to those who took part in the study. How do these principles impact our choice of the level of randomization?

We need to be careful to ensure that randomizing at an individual level does not create tensions that lead to harm if some in a community are seen as being unfairly chosen for benefits. If the process of randomization is not properly explained and accepted, having a few individuals in a community suddenly receiving large benefits compared to others might make them targets for criticism or worse.

Another key aspect of ethical behavior is ensuring that participants provide informed consent. (Module 2.4 has further details on different ways to seek informed consent that are appropriate to studies with varying levels of risk.) Randomizing at the group level raises difficult issues about seeking informed consent. Normally we explain the study and ask for consent when we interview an individual, but with group-level randomization we often do not interview all those who receive the program. Whether those who given the program but are not part of the data collection are part of the study and thus whether we need to ask their consent depends on the situation (definitions of study participants also vary by IRB). For example, it may depend on whether participation in the program is voluntary, how involved the evaluators are in the program design, and the level of risk. Usually evaluators seek to inform the community as a whole (through a community meeting) and get “community-level consent.” We strongly advise discussing the appropriate approach with the relevant IRB regulator.

#### Politics and feasibility

*Is the program allocation perceived as fair?* Allocation is likely to seem arbitrary when people with equal needs who interact on a regular

basis are assigned to different groups. To avoid the appearances of unfairness, it may be necessary to randomize at a higher level, such as the community. Randomizing at the community level helps partly because those in the treatment and comparison groups may not interact as regularly. In developing countries, at least, it is often considered normal for some communities to receive benefits (such as a well or a school from an NGO or a government agency) that others do not receive. This may be less true in developed countries where communities are used to national-level policies' dictating the allocation of programs.

*Is it permitted?* Sometimes the authorities may deny permission to randomize at the technically conducive level. For example, we may want to randomize across schools, but the authorities want every school to receive the program. This may require a design in which every school has one grade that benefits from the program but which grade benefits is randomized.

#### Logistics and feasibility

It is usually infeasible to randomize at a lower level than that at which the program is administered. If a program builds a new market center for a community, it is impractical to exclude some members of the community from purchasing produce there. Even when specific services are delivered to individuals, it is still sometimes easier to randomize across groupings of individuals (clusters) than across the individuals: delivering school meals to some children and not to others within a school would be logistically difficult because we would have to separately identify eligible children.

#### Cost and feasibility

If we randomize at the village level, we usually have to treat everyone in a village. Does the partner organization have the money to deliver the intervention at that level? Do we have the money to perform an evaluation at this level? Village-level randomizations tend to have much higher transport costs than those done at the individual level.

#### **Reality check of units of randomization**

Often the common units of randomization we mention above do not exist as neatly in real life as they do in textbooks. Official village boundary lines in government records may not correspond with the actual boundaries according to which people live and interact. Sepa-

rate school classes may exist on paper, but in reality they may often be merged because of a lack of teachers or because teachers are chronically absent. For example, in Bangladesh the population density in rural areas is often so high that one “village” merges into the next, with no real distinction observable on the ground. In Kenya’s Western Province, people live not in village clusters but on their farms, which are evenly spread out, making it difficult to randomize by “community.” Administrative distinctions are not a good proxy for close social interaction in these cases, and we may have to come up with our own definition of the cluster (grouping of individual units) at which we want to randomize. We may want to use participatory methods to understand what local people consider their community. Alternatively, we can randomize using units not directly related to our program. For example, an agricultural program in the Western Province of Kenya used the primary school catchment area as its unit of randomization.

---

#### **MODULE 4.3 Deciding Which Aspects of the Program to Randomize**

*We can randomize three aspects of the program: access to the program, the timing of access, and encouragement to access the program. In this module we discuss five research designs that we can use to create randomized variation in exposure to the program: the treatment lottery, the lottery around the cutoff, the phase-in design, the rotation design, and the encouragement design. This module reviews these four designs and a variation of the treatment lottery. The pros, cons, and ethical considerations of each are discussed.*

##### ***The treatment lottery***

*Allocation.* The units (individuals, households, schools, etc.) are randomly assigned to the treatment and comparison groups. The treatment group is offered access to the program; the comparison group is not. The two groups keep this status for the duration of the evaluation.

*Difference in exposure to the program.* This comes from offering access to the program to the treatment group but not to the comparison group.

When is the treatment lottery most workable?

*When the program is limited in scale or being piloted.* It is easier to justify a randomized evaluation when there are only enough resources to

provide access to a small number of people or when the program is untested and the effects unclear.

*When the program is oversubscribed.* The lottery provides a conspicuously fair allocation mechanism.

*When the evaluation will measure effects in the long run.* The lottery allows us to measure the long-term effects of the program because there is always a comparison group that never receives the program.

What to consider when using the treatment lottery

*Only in special circumstances can treatment lotteries be used to evaluate entitlement programs.* Entitlements are programs that people receive as a legal right. Usually government regulations do not allow people to be randomized into and out of eligibility for entitlement programs. However, some countries have made exceptions to allow for the evaluation of their entitlement programs. In the United States, federal guidelines establish rules for federally supported entitlement programs (like Medicare). States are allowed to request waivers to deviate from these guidelines.<sup>18</sup> Waivers of welfare policy are conditioned on running a rigorous (usually randomized) evaluation of the revised program.<sup>19</sup> In these circumstances, individuals in a single state would experience different program benefits and conditions.

For example, in most US states, welfare benefits are based in part on family size. In the early 1990s, policymakers were concerned that giving more assistance to larger families encouraged women to have more children than they would have had otherwise. Several states applied for federal waivers to test new policies that would place a cap on the additional benefits families could receive if they had more children while on welfare. New Jersey was the first state to randomly assign a family cap to a subset of its welfare beneficiaries in order to evaluate its impact on childbearing beginning in 1992.<sup>20</sup>

18. Social Security Act. 42 USC 1315 §1115. [http://www.ssa.gov/OP\\_Home/ssact/title11/1115.htm](http://www.ssa.gov/OP_Home/ssact/title11/1115.htm).

19. Prior to US welfare reform in 1996, waivers for states to redesign aspects of federal entitlement programs were conditioned on evaluating the new program. The Personal Responsibility and Work Opportunity Reconciliation Act of 1996 specified randomized evaluation as the preferred methodology. See the Office of Family Assistance 2012 information memorandum "Guidance Concerning Waiver and Expenditure Authority under Section 1115," TANF-ACF-IM-2012-03, <http://www.acf.hhs.gov/programs/ofa/resource/policy/imofa/2012/im201203/im201203?page=all>.

20. US General Accounting Office, "Welfare Reform: More Research Needed on TANF Family Caps and Other Policies for Reducing Out-of-Wedlock Births," GAO-01-924, 2001, <http://www.gao.gov/new.items/d01924.pdf>.

*Attrition levels may be higher in lottery designs.* Sometimes people drop out because they were not assigned their preferred treatment. In a treatment lottery, those in the comparison group will not receive the program in the future, and thus attrition may be more of a problem.

#### Ethical considerations in using the treatment lottery

Unlike in the case of some other methods of randomization, in a treatment lottery some participants in the study are never given access to a program. Is this ethical? The treatment lottery approach is often used when there are insufficient funds to provide the policy or program to all those who could benefit from it. For example, an NGO program may target scholarships to cover four years of secondary education but not have enough funds to cover all eligible children. We may use a lottery to decide who receives access to the program. Recalling the ethical principles described in Module 2.4, as researchers we must carefully weigh the benefits of the evaluation against the risk of harm and seek to minimize potential risks. Because the evaluation has not changed the number of beneficiaries of the program, we have not changed the risk–benefit profile. The only exception is if harm is caused by a participant seeing someone else’s benefit while not benefiting herself. In this case we may need to think about randomizing at a higher level, as discussed above in the section on the unit of randomization.

But what if there is sufficient funding to cover all eligible participants? Is it ethical to use the treatment lottery in this case, when the lottery will reduce the total number of beneficiaries of the project, at least in the short run? Ethical guidelines suggest that we must trade off the potential harm of treating fewer beneficiaries as a result of doing an evaluation against the potential benefits of the evaluation. How can we do this? We need to take into account the level of uncertainty about the benefits of the program, how many people will have delayed access to the program because of the evaluation, and what the benefits would be of having more information about the impact of the program.

It can be ethical to proceed with an evaluation if any of the following conditions hold: if it is unclear whether the program will have positive effects, if there is some risk that the program will have negative effects, or if there are likely to be benefits to society in general and to one group (say, girls) in particular from learning the impact of the program. (Remember that the justice consideration means that bene-

fits should accrue to the group in society that is undertaking the risks, and thus it is important that the benefits of doing the evaluation accrue to orphans in this example—not just to society in general.)

The benefits of undertaking the research may include being able to use the evidence of program effectiveness to raise more funding to expand the program so that it reaches more people in the long run, the gain from replacing the program with something more effective if we find its impact is disappointing, and the benefit of avoiding harm if we find that the program has unintended negative consequences and it is not scaled up as originally planned.

*Lotteries and medical ethics.* Medical trials have particular ethics rules that must be followed when using a lottery design. Under these rules, it is ethical to do an experiment to test two alternative treatments only if, at the start of the trial, we do not know whether the existing treatment or the new treatment is better. However, once it is clear that one approach is better than the other, there is an ethical obligation to end the trial early and scale up the effective treatment to all those in the evaluation.

In practice, this issue rarely arises in randomized evaluations of social programs. In medicine there is a presumption that everyone is receiving care and the question is which type of care is most effective. In most social programs, only some of those who could potentially benefit are receiving any program at all. Even if we know what is effective, there is often not sufficient funding to cover everyone. Additionally, we rarely have enough data to know for sure that a program is effective partway through the evaluation. In those cases in which we do have a lot of data in the middle of the program, we may not have the money to scale it up to everyone if we find midway that the program is succeeding, but we may want to consider stopping the program early if we find that it is having a negative effect.

### ***The treatment lottery around a cutoff***

*Allocation.* For programs that select participants based on qualifications (e.g., credit programs, secondary schools, need-based programs), a regular treatment lottery may not be feasible. Banks, for example, may not be willing to lend to any “random” applicant but may want to lend to all well-qualified applicants. Randomized evaluation, however, is still possible if we divide applicants into three groups: those who will be accepted into the program no matter what, those who will not be accepted into the program, and those who will have a random

chance of being accepted into the program. Which group people fall into depends on how well qualified they are for the program.

Examples of groups that could be randomized around the cutoff include students seeking scholarships to go to university, applicants for loans, and families with assets slightly too great to qualify for an existing welfare program.

There are three slightly different versions of a lottery around the cutoff:

1. *A lottery among the marginally ineligible.* The program accepts all those it would have previously accepted, but now it is expanded to also include some who had previously just missed the cutoff. This tests the impact of expanding eligibility to the program—for example, expanding access to secondary education or expanding access to subsidized health insurance to those just above the poverty line. This design is possible only with additional program funding
2. *A lottery among the marginally eligible and marginally ineligible.* If there are no resources to fund additional participants, it is still possible to use the lottery around the cutoff design. Some of those who would previously have been eligible are put into the lottery along with some of those who would previously have fallen just below the cutoff. This design works best when new recruits are being accepted into the program. It is very hard to use a lottery to remove some existing participants from a program and at the same time accept others into the program. This design tests the impact of the program at the margin—both the impact of expanding and that of slightly contracting the program.
3. *A lottery among the qualified.* In this design there are only two groups, the eligible and the ineligible. The lottery takes place among all those who are eligible, but we still have a group that is entirely rejected. The only requirement is that we have more eligible people than we have places for. The advantage of this approach is that we are now testing the impact of the program on the average participant. This design is very close to the simple lottery described above.

*Difference in exposure to the program.* This comes from randomizing access to the program among a particular subset of those who apply.

What to consider when using the lottery around a cutoff

*Does this design answer a relevant policy question?* A lottery around the cutoff tests the effect of the program on a very specific population—those who are close to the cutoff. For example, the lottery among the ineligible tests only the impact of expanding the program to those not currently eligible. There is no comparison group of those currently eligible, because they all remain eligible. Sometimes whether there is a comparison group is exactly the question we want to answer because the policy question before us is whether to expand the program to those who are currently not eligible. For example, an evaluation in Oregon is examining the effect of making Medicaid available to those just above the current cutoff for Medicaid eligibility. Whether to expand health insurance to those who are poor but currently do not have access to medical insurance is a major policy question in the US currently. However, the evaluation will not be able to say what impact Medicaid will have because there is no comparison group among those currently eligible for Medicaid.

*How large a range around the cutoff should we use?* Imagine that we have 200 applicants for 100 places. We rank all the applicants from 1 to 200. Normally the program would accept the first 100 and reject the rest. For our evaluation, should we accept the first 50, reject the bottom 50, and randomize acceptance among the middle 100? Or should we accept the top 80, reject the bottom 40, and randomize among the remaining 80? The answer depends on program needs, perceived fairness, and statistical power. The program may have some basic requirements for its participants; those who do not fulfill these should be rejected.

We discuss ethical considerations and perceived fairness (political considerations) in more detail below. But it is worth noting that sometimes considerations around perceived fairness and statistical power may push us in different directions. In some situations it may be seen as unfair if some people are accepted into the program who score much lower on the qualification criteria than others who are rejected. For this reason, there may be pressure to keep the range of people from which we randomize quite narrow: for example, we could randomize among those who scored between 50 and 55 on the qualifying criteria. This difference is likely to appear small to most people, and it should be easy to convince the community that it is pretty arbitrary whether someone got a score of 50 versus 55, so randomizing



participation is fair. But if we have a narrow range, this restricts the sample size we can use to estimate the impact of the program and thus the statistical power of measuring the impact. There is little point in performing the evaluation if our sample size is very small (see Chapter 6).

One way to try to balance these two competing pressures is to look for a point at which there are lots of people within a small range of scores and make that the randomization range. Then we can have a reasonable sample size and still randomize over a relatively small range of scores. As discussed above, our evaluation will test the impact of the program only on those whose scores fall within this range.

#### Ethical and political considerations

Usually when we introduce a lottery around the cutoff we do not change the number of beneficiaries, but we do change who receives access to the program. Is it ethical to deny access to those who are more qualified and give their places to people who are less qualified for the sake of the evaluation?

In assessing the trade-off between the costs and benefits of using a lottery around the cutoff, there are a number of issues to keep in mind. First, we are unlikely to know for sure that having access to the program is beneficial or we would not be doing the evaluation. As we discussed previously, there are degrees of uncertainty: the stronger the evidence that the program is beneficial, the more we have to be concerned about the harm of “denying” people access. In the case of a lottery around the cutoff, because we are changing the types of people who receive the program, a key question is whether we know that the benefits of the program are likely to be greater for those who are more qualified.

For example, imagine that we are evaluating the effect of giving access to consumer loans to people in South Africa.<sup>21</sup> The bank has a scoring system it uses to decide which applicants are creditworthy. The assumption is that those who score highly will be able to repay the loans, whereas those who score badly will not be able to repay, potentially getting themselves into a damaging debt trap. Both the bank and the participants are better off with scoring. But do we know that

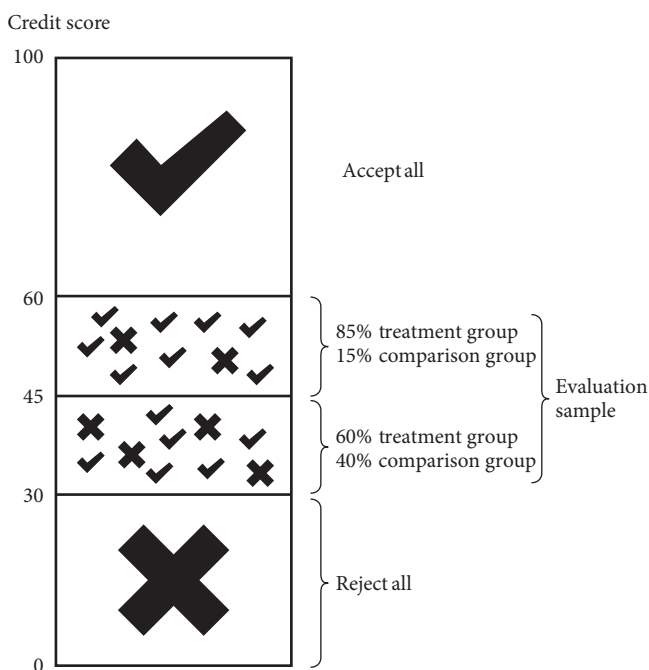
21. This example is inspired by but is a simplified version of Dean Karlan and Jonathan Zinman, “Expanding Credit Access: Using Randomized Supply Decisions to Estimate the Impacts,” *Review of Financial Studies* 23 (2010): 433–464.

the scoring system is good at determining who is a good risk and who is a bad risk? Maybe the system is good enough to detect the very good risks and the very bad risks, but does it do a good job of selecting people around the cutoff? The credit scoring system may be discriminating against people who are otherwise good risks but happen to receive a low score on some characteristic, such as living in a poorer neighborhood.

If there is uncertainty about the quality of the scoring system, a lottery around the cutoff can help us learn how good the system is and whether the cutoff has been placed at the right point. If we find that those just below the cutoff do just as well as those above it, then the bank will be encouraged to extend its loans to more people and those just below the cutoff will gain, as will the bank. There is a risk that the cutoff was at the right place initially and that those below the cutoff will get into debt as a result of being offered a loan they cannot repay. We have to take this risk into account when designing the study. We can ameliorate this risk by randomizing only above the cutoff (a lottery among the qualified), but this has other issues: we don't learn whether the cutoff was too high, and we reduce access among the qualified more than in other designs.

In general, the better the evidence we have that the cutoff is well measured and targets the program well, the more careful we need to be with a lottery around the cutoff. For example, researchers decided not to evaluate a feeding program for malnourished children because the criteria for selecting participants (height for weight and arm circumference) were relatively well designed. In contrast, credit score cutoffs used to gauge the creditworthiness of the poor clients served by MFIs do not have a lot of research behind them, and this level of uncertainty increases the benefits and reduces the probability of harm of changing the cutoffs.

One way to make a lottery around the cutoff more politically acceptable is to vary the probability that access to the program will depend on people's scores. For example, we could give those above the cutoff a higher (though still random) probability of being accepted into the program than we give to those below the cutoff. We call the probability of being assigned to treatment the *allocation fraction*, so in these examples we are varying the allocation fraction depending on the score. Figure 4.4 provides an illustration. In our example of randomizing access to credit around a credit score cutoff, we accept all those with a credit score above 60 and reject all those with a credit



**FIGURE 4.4** Lottery around the cutoff with varying allocation fractions

Note: Checkmarks indicate acceptance; x's indicate rejection.

score below 30. Those with a credit score between 45 and 69 have an 85 percent probability of being given credit, while those with a credit score between 30 and 45 have a 60 percent chance of receiving credit. In this design we have included an element of randomization, but each individual's probability of receiving credit is closely linked to her credit score.

### ***The phase-in design***

*Allocation.* When everyone must receive the program, we can randomly select who is phased in to the program first and who receives it later.

*Difference in exposure to the program.* The groups not yet phased in form the comparison group. They become the treatment group when they are phased in.

When is the phase-in most workable?

*When everyone must be given access eventually.* When everyone must receive program services during the evaluation period, the phase-in design allows us to still create a temporary comparison group.

*When not everyone can be enrolled at once.* When logistical or financial constraints mean that coverage must expand slowly, randomization may be the fairest way to choose who receives the program first. Programs with a heavy upfront training component will often fall into this category. If we are using a phase-in design because of financial constraints, we must be sure that we will have the funding and other resources to deliver the program later as we promise.

*When anticipation of treatment is unlikely to change the behavior of the comparison group.* People assigned to the later phases of a program could make choices based on expected treatment. Imagine a phase-in program providing capital grants of US\$100. People know that next year they will receive a grant of US\$100. They can borrow against that and have the capital this year, undermining our counterfactual. This is more likely to happen if the program is providing high-value transferable goods and services. If we are providing such untransferable goods as deworming pills to be swallowed on the spot, anticipation effects are unlikely.

*When we are interested in the average program impact over different years.* Often when we use a phase-in design we end up pooling the data across the several years of the experiment. However, with enough sample, we could look at each year separately and ask, “What was the impact of the program in year 1, when Group A was the treatment group and Groups B and C were the comparison group?” Then we ask, “What was the one-year impact of the program when Group B had had the program for one year and Group C was the comparison group?” Finally, we can ask, “What was the two-year impact of the program when Group A had had the program for two years versus Group C, which was the control?”

Usually we do not have enough statistical power to ask these three separate questions. Instead we perform one test that uses all our data. We implicitly assume that the effect on A in the first year of the evaluation is the same as the effect on B in its first year of the program but the second year of the evaluation. Alternatively, we may accept that the impact may vary in different calendar years but be satisfied with

calculating the average effect of the program over the two calendar years. Note that in the phase-in setup we can account for changes in the general environment in different calendar years (for example, farmers do worse in the second year because of a drought) because we have our comparison group, Group C, which we use to factor out common-year effects. We can also pick up whether the program had more of an effect in its second year than in its first by comparing Group A in the second calendar year to Group B in the second calendar year (although we may not have much power to pick up small effects).

What to consider before using the phase-in design

*Attrition may occur.* Because it promises goods and services in the future, the phase-in design may increase the chance that the comparison group will continue to cooperate with the follow-up surveys and tests, which limits attrition.

*Anticipation of treatment may induce behaviors that could mask the effects of treatment.* As discussed above, anticipation of program goods and services can change the current behavior of those yet to be phased in, which can bias the impact estimates. Anticipatory effects can go in different directions: if the prospect of being given a free bed net makes people less likely to buy one now, the evaluation will overestimate the impact of the program. If anticipation of a grant tomorrow makes people more likely to invest today, the evaluation will underestimate the impact of the program. Anticipatory effects are one of the biggest drawbacks of the phase-in design. It means that this design is most workable when evaluating programs in which these effects are unlikely to exist. In some cases, program implementers do not tell individuals or communities that they will be phased into the program at a later stage because of concerns that the rollout may be prevented for some reason. This helps dampen possible anticipation effects, although it also reduces one advantage of phase-in programs, that communities will participate because they know they will receive the program later.

*The time to effect (changes in outcomes) must be shorter than the time to the last phase-in.* With each group phased in, the treatment group grows and the comparison group shrinks. If the phases are too short, there may not be enough time for the program to have a measurable effect. Imagine that we have three groups and are phasing in new participants every six months. All the groups will be phased in by the end of the first year. If the program takes two years to achieve its effect, the evaluation will find no impact. A well-articulated theory of

change and logical framework, preliminary data collection, historical data, and existing literature can all help in guessing the time needed to achieve an impact.

Long-run effects cannot be measured unless the program targets a specific age or cohort, in which case some people will never receive the treatment. Once everyone is phased in, there no longer is a comparison group. The long-run effects cannot be estimated. One exception arises when people age out of the program before they are phased in. An example is education programs. If we are treating a given cohort, say children in the last year of school, every year the graduating cohort is aged out. If a school is randomly assigned to be phased in during the third year, two cohorts will age out before they receive the program. These children will never receive the program and can be used as a long-run comparison group.

*Ethical considerations for phase-in designs.* Because everyone is offered the program eventually, the ethical issues raised by the phase-in design center on the time to treatment. Even when they have enough funding, implementing organizations often face capacity constraints that force them to roll out programs slowly in the absence of an evaluation. For example, there may be enough funds to deworm all children in a district but not enough program staff to train all the teachers to provide the deworming medicine. In the absence of evaluation, the program will be rolled out based on the logistical preferences of the implementers: for example, those who live nearest the implementers' headquarters may receive the program first. What the phase-in design does is to formalize the rollout into rigid phases and randomize who is included in each phase. Implementers may face some costs associated with having the initial phases more geographically dispersed that would not have been incurred otherwise. These costs need to be weighed against the benefit associated with the evaluation.

*A phase-in design can still be ethical even when there are sufficient financial and logistical resources to treat everyone immediately if the benefits of learning the impacts of the program are sufficiently high. In either case, the ethical principles described in Module 2.4 apply: we must carefully weigh the benefits of the evaluation against the risks of harm and seek to minimize potential risks.*

### **The rotation design**

*Allocation.* When everyone needs to receive a program but resources are too limited to provide access to everyone at once, we can rotate

who receives the program. We can randomly divide those who are to receive the program into two groups, and the two groups can take turns receiving the program. The group that receives access to the program forms the treatment group for that period, and the other group forms the comparison group. Once it is the comparison group's turn to receive the program, it becomes the new treatment group, and the earlier treatment group switches to be the comparison group. We can randomly select the order in which the groups take turns.

*Difference in exposure to the program.* This comes from offering the program to one group and withholding it from the other at any given time.

When is rotation design most workable?

*When resources are limited and are not expected to increase.* When resources will remain limited for the duration of the evaluation and everyone has to be treated, the participants will have to take turns. The same applies when burdens have to be shared fairly.

*When the primary concern is what happens to people during the time they have access to the program.* For example, how do people behave when they are in office (and have official powers) or when they have access to insurance? In these cases, the potential effects of the program on behavior may go away once these individuals no longer have the program.

*When the treatment effects do not remain after the treatment ends.* Because the original treatment group switches and becomes the comparison group once they are out of it, we must be sure that the program has only short-term effects. If there are lingering effects, the original treatment group will not make a valid comparison group once their turn is over because their outcomes will reflect whatever program effects remain. These lingering effects will distort our impact estimate.

*When we want to measure or document existing and induced "seasonal" influences.* Rotation induces seasons: seasons when people have access to the program and seasons when they do not. But for some programs, there already are existing seasons. For example, schools can be in session or on vacation. Agricultural programs or antimalarial interventions are affected by the rainy seasons. How does this seasonality influence program outcomes? When and how is a seasonal program best delivered? How do the existing seasons and the "seasons" introduced by the rotating program interact? Because a rotation

design introduces random variation as to when people are exposed to the program, it is possible to use it to measure the best time to run the program.

Imagine a tutoring program that is run both after school and during school vacations. Every three months children are rotated into and out of the program. Some children will experience the program as an afterschool program, and others will experience it during school vacation. At which time will they learn more—during the term, when the tutoring complements what they are doing in school, or during vacation, when their learning would perhaps decay rapidly given that they would not be doing any other schoolwork?

What to consider before using the rotation design

*Rotation is common in everyday life and easy to understand.* Here are some examples:

- *Vacation timeshares* People take turns occupying a jointly owned vacation home.
- *Rolling blackouts or load shedding* When electricity or water is being rationed in a city, the neighborhoods take turns having water and electricity at favorable times, such as rush hours.
- *Afternoon school and night school* When there are more classes than classrooms, rotation is often used to share the classrooms.
- *Rotating savings and credit associations* Members take turns using a collective savings pool.

*Usually there is no pure comparison in the long run.* Every group is treated at some point, so we cannot measure long-term impacts. The exception is if we rerandomize.

*Rerandomizing can allow us to measure the effects of length of exposure to a program.* The usual way to rotate the treatment is that groups take alternating turns repeatedly. For example, if we had two groups, A and B, the alternations would create one of the following treatment patterns: A, B, A, B or B, A, B, A. Instead we can rerandomize for each period. Imagine flipping a coin at the beginning of each time period to decide which group receives the treatment for that period. Sometimes the same group may receive the treatment again. For example, the repeated coin toss may create this pattern: A, B, B, A. If we compare the outcomes of Group B after two consecutive turns to those



during the other periods, we may be able to disentangle the effects of length of exposure. For example, the political quotas for women in some states in India are rerandomized at each election, which creates variation in the length of time voters are exposed to women in political leadership.

*Anticipation of having, or not having, treatment may change present behavior.* As in the case of phase-in, the out-of-treatment group may behave differently in anticipation of receiving the treatment, and the in-treatment group may anticipate not having the treatment, which can undermine the validity of the comparison group.

*The time needed to change outcomes may be shorter than the treatment period.* The lag between the time of treatment and the time that the effect becomes detectable should be shorter than the treatment period. If the program rotates before then, impact estimates may be distorted.

*The program must affect only those currently in treatment.* Unless there are no lingering effects, we cannot have a “pure” comparison group. The exception is if we rerandomize after each cycle and by chance we have a group that is never treated. However, this would undermine one of the main rationales for using the rotation design: that everyone receives access to the program.

*Ethical considerations.* Ethical considerations for the rotation design are similar to those for the phase-in design. The difference is that in rotation people exit the program. It is therefore important to consider whether there are any risks associated with having a program withdrawn that are different from those of not having exposure to the program. For example, providing a short-term subsidy to address a long-term problem might be harmful. We would not want to provide antibiotics for only part of the time someone was sick. We might not want to provide a loan subsidy for only part of the time someone will have a loan because it might encourage people to take on unsustainable debt.

### ***The encouragement design***

*Allocation.* When a program is open to all, access cannot be directly randomized, but there is still an opportunity for evaluation if the program has low take-up. A treatment group is created by randomly assigning individuals or groups to receive encouragement to take up the program. The encouragement can be in the form of a postcard reminding people of the program, a phone call, a letter, or a reward.

The idea is to increase the probability of take-up by the encouraged. The comparison group still has access to the program but does not receive special encouragement to take it up.<sup>22</sup>

*Difference in exposure to the program.* Anyone who comes forward will receive treatment, regardless of whether they received the encouragement or not. Difference in exposure at the group level is created when a higher proportion of those in the treatment group take up the program than do those in the comparison group.

When is the encouragement design most workable?

*When the program is open to all and is undersubscribed.* The design works best when there are enough resources to treat everyone but take-up is low. This low take-up creates the opportunity to create higher take-up in the treatment group.

*When the program is open to all but the application process takes time and effort.* When the application process is burdensome, we can offer help with the application to some but not to others.

*When we can find an encouragement that increases take-up but does not directly affect outcomes.* The encouragement research strategy works only if the encouragement does not itself affect the behaviors and outcomes the evaluation is targeting and measuring. The challenge is to use a form of encouragement that creates a big difference in take-up rates between treatment and comparison groups without directly affecting outcomes. Here the encouragement is acting as an instrument in predicting take-up. In other words, this is a randomized form of the instrumental variables strategy we discuss in Module 2.2. The assumption that the encouragement affects outcomes only through its effect on take-up is exactly the same as the assumption for any instrumental variables strategy and is called the exclusion restriction. The benefit here is that we know the instrument itself is randomly allocated.<sup>23</sup>

Imagine that we are interested in the effect of a training program for pepper farmers that is taught in a local market town and is designed to

22. An early example of the encouragement design is Marvin Zelen's "A New Design for Randomized Clinical Trials," *New England Journal of Medicine* 300 (1979): 1242–1245.

23. For a more detailed description of the "exclusion restriction," see Joshua Angrist, Guido Imbens, and Donald B. Rubin, "Identification of Causal Effects Using Instrumental Variables (with Discussion)," *Journal of the American Statistical Association* 91 (1996): 444–472.

increase the quality of the peppers grown by farmers. We find that many local pepper farmers do not take up the training, and one of the reasons given is that they cannot afford the bus fare into the market town to attend the training. We therefore think about offering free bus passes to a random sample of pepper farmers as an encouragement to attend the class. We will measure the success of the program by examining the price at which the farmers sell their peppers (the theory is that the training helped increase the quality of their peppers and that higher-quality peppers receive higher prices). But our approach falls afoul of the exclusion restriction. The problem is that giving some farmers free unlimited transport to the local market town may affect the prices at which farmers sell their peppers even if they never attend the training or they learn nothing from it. Farmers without the transport subsidy may sell their peppers to middlemen who bear the transport costs of getting the peppers to market, while those with the free transport subsidy may sell their peppers in the local market town and receive higher prices because they cut out the middlemen. Our encouragement (the transport subsidy) may have increased take-up of the training, but it has also influenced our outcome measure (the sale prices of peppers) through a mechanism (place of sale) other than the effect of the program (increased training in improving pepper quality).

Often encouragement mechanisms involve some form of subsidy to take up the program, such as a small incentive tied to take-up. It is important that any subsidy be very small; otherwise the monetary benefit of the encouragement may have a direct effect on outcomes by making people richer. The most workable encouragement designs are specifically targeted to the take-up of the program.

It may seem hard to find an encouragement that is small enough not to have a direct effect on the outcome and yet have a large effect on take-up. But behavioral economics is increasingly finding that relatively small nudges can have surprisingly large effects on behavior. For example, helping people fill in the application form for a program may seem like a small incentive. If the service is valuable, surely people would fill out a form without help, we reason. But a number of studies have found that help with applications can have a large impact on program take-up.<sup>24</sup>

24. See, for example, Céline Braconnier, Jean-Yves Dormagen, and Vincent Pons, "Willing to Vote, but Disenfranchised by a Costly Registration Process: Evidence from a Randomized Experiment in France," APSA 2012 Annual Meeting Paper, <http://ssrn>

What to consider before using the encouragement design

We have already set out a number of issues to consider when using the encouragement design while discussing the situations in which an encouragement design is most useful. Here we discuss two others.

*Those who respond to the encouragement are a policy-relevant group.*

An encouragement design measures the impact of the program on those who respond to the encouragement. It does not tell us the impact of the program on those who had already taken up the program, nor does it tell us the impact on those who did not respond to the encouragement. Before we use an encouragement design, we should think carefully about the question we want to answer. Do we want to measure the impact of the program on the average person who takes it up, or do we want to measure the impact of the program on the marginal person who takes it up? If we want to measure the average impact across all those who take up the program, we want those who respond to the incentive to be comparable to those who have already taken up the program. But an encouragement design is particularly useful when we are especially interested in the impact on those who respond to the encouragement (were not taking up the program before). For example, if we want to know whether to invest in encouraging more people to take up the US food stamps program, we will want to know the impact of taking up food stamps on those who respond to an encouragement program.

*The encouragement must not encourage some and discourage others.*

When we use an encouragement design, we must make what is called the *monotonicity assumption*. This means that everyone must be affected by the encouragement in the same direction. If we provide an incentive to take up a program, it must either increase the probability that someone will take up the program or not impact anyone at all. If the encouragement itself increases the take-up of some groups and reduces the take-up of others, we will likely get a biased estimate of the impact of the program. All too often, researchers play little attention to this assumption because it seems natural to assume that an incentive or an encouragement will have only a positive effect on take-

---

.com/abstract=2108999, and Florencia Devoto, Esther Duflo, Pascaline Dupas, William Pariente, and Vincent Pons, "Happiness on Tap: Piped Water Adoption in Urban Morocco," NBER Working Paper 16933, National Bureau of Economic Research, Cambridge, MA, 2011.

up. But there are some cases, like that of the provision of information, in which both positive and negative responses may be quite natural. For example, if we tell people the benefits of staying in school, in terms of increased earnings, this is likely to have different impacts on whether children stay in school depending on whether they had previously under- or overestimated the likely benefits of staying in school. The information may lead some to stay in school longer, while it may lead others to stay less long.

In Module 7.1 we explain why these “defiers” can undermine the validity of our experiment. For now, however, it is important to note that we need to carefully think through the channels through which an encouragement might work and make sure that it will not have perverse effects on (discourage) some people.

#### Ethical considerations

In some ways, encouragement designs ease some of the ethical and political concerns that may arise when randomized evaluations restrict access to certain programs. Everyone can have access to the program; we simply make it easier for some to have access than for others. But this difference is really a matter of degree. If we make it easier for some people to gain access to the program than for others we are still potentially (if the program works) giving a benefit to some over others, so we should consider whether the evaluation is changing those who access the program in a way that could be detrimental. In other words, we still need to go through the process of balancing the risks and benefits of the evaluation even in the case of an encouragement design.

#### *Hybrid strategies in multistage experiments*

Some research questions may require us to randomize more than once. We can use these multistage experiments to evaluate different components of an intervention or concepts underlying behavior. When we perform multistage randomization, we can use different randomization strategies at each stage to create hybrid strategies.

#### Hybrids when isolating spillover effects

As we discuss in Module 4.2, we may want to randomize in two stages to measure social interaction or spillover effects. For instance, researchers want to examine the social effects and information effects of participating in a retirement savings program offered by a large employer.

The employer offers information sessions about the program to employees. The organization is divided into many departments. We can use a combination of the treatment lottery and the encouragement design to generate variation in treatment density. We can randomize the proportion of employees in each department who will receive extra encouragement to sign up for the information session and then randomize which individuals in each department receive the encouragement.

Hybrids to isolate channels of underlying behavior

When there are multiple possible impact channels and we want to separately identify each, we may have to randomize more than once. An experiment in South Africa used a two-stage approach to isolate moral hazard and adverse selection, two concepts thought to explain why it is so hard to lend to the poor. In the first stage, clients of a South African lender received letters offering loans with randomly assigned high and low interest rates. This stage used the encouragement design. Those responding to low-rate offers were offered low-rate loans. (We call them the low-to-low group, because their repayment burden was low and remained low.) Those responding to high-rate offers were randomly split into two groups. This stage used the lottery. Half were randomly surprised with a lower-rate loan (the high-to-low group), and the other half were offered the original high rate (the high-to-high group).<sup>25</sup> The theory of moral hazard says that borrowers with little at stake face great temptation to default if the repayment burden becomes too large. If there is moral hazard, the clients who borrow at the higher rate are more likely to default because they have a large repayment burden. Decreasing the repayment burden for some reduces the chances that they will be subject to *moral hazard*. Comparing the high-to-high and the high-to-low groups identifies moral hazard. *Adverse selection* essentially says that, since the interest rate is supposed to reflect the risk, the less risky clients will refuse to borrow at high interest. If there is adverse selection, clients who agree to borrow at a higher rate are more likely to default. Comparing the high-to-low and the low-to-low groups identifies adverse selection.

25. Dean Karlan and Jonathan Zinman, "Observing Unobservables: Identifying Information Asymmetries with a Consumer Credit Field Experiment," *Econometrica* 77 (2009): 1993–2008.

### Module 4.3 summary: Comparing the four basic strategies

Strategy	Randomization design	Most workable when	Advantages	Disadvantages
Basic lottery	<ul style="list-style-type: none"> <li>• Randomizes access to program</li> <li>• Leaves treatment status unchanged throughout</li> <li>• Compares those with and without access to the program</li> </ul>	<ul style="list-style-type: none"> <li>• The program is oversubscribed</li> <li>• Resources are constant for the evaluation period</li> <li>• It is acceptable that some receive no program</li> </ul>	<ul style="list-style-type: none"> <li>• Familiar and well understood</li> <li>• Usually seen as fair</li> <li>• Easy to implement</li> <li>• Allows for estimation of long-term impacts</li> </ul>	<ul style="list-style-type: none"> <li>• Differential attrition—units in the comparison group may have little reason to cooperate with the survey</li> </ul>
Lottery around the cutoff	<ul style="list-style-type: none"> <li>• Randomizes access to the program among those close to an eligibility cutoff</li> </ul>	<ul style="list-style-type: none"> <li>• The program determines eligibility with a scoring system</li> <li>• There are large numbers of participants</li> </ul>	<ul style="list-style-type: none"> <li>• Program has considerable flexibility about who to enroll</li> </ul>	<ul style="list-style-type: none"> <li>• Measures the impact of the program only on those who are close to the eligibility cutoff</li> </ul>
Phase-in	<ul style="list-style-type: none"> <li>• Randomizes the timing of access</li> <li>• Switches units from comparison to treatment over time</li> <li>• Compares those with access to those who have not yet received access</li> </ul>	<ul style="list-style-type: none"> <li>• Everyone must eventually receive the program</li> <li>• Resources are growing over time</li> <li>• Treatment is being replicated</li> </ul>	<ul style="list-style-type: none"> <li>• Common</li> <li>• Easy to understand</li> <li>• The comparison group is more likely to cooperate in anticipation of future benefits</li> </ul>	<ul style="list-style-type: none"> <li>• The comparison group eventually goes away</li> <li>• Anticipation of treatment may affect the behavior of the control group</li> <li>• There is a limited time over which impact can be measured</li> </ul>

### MODULE 4.4 The Mechanics of Simple Randomization

*This module outlines the mechanics of random assignment. We discuss the ingredients we need, the steps involved, and the pros and cons of different devices we can use to perform the randomization.*

### *The ingredients of random assignment*

We need five ingredients to perform random assignment:

1. A list of eligible units (such as people, communities, schools)
2. The number of randomization cells
3. Allocation fractions
4. A randomization device
5. Initial data on the eligible units (for stratification or balance check)

#### The list of eligible units

The list of units eligible for our program can be individuals or groupings of individuals, such as households, villages, districts, schools, or firms. In the illustrations below, we consider the 26 letters of the alphabet as our list of eligible units (our *sampling frame*). In real life we have to create that list somehow. Here we provide some examples of commonly used methods for acquiring this list and sources of information that can be used.

#### *Existing data from the government and other providers*

*Local governments.* Local authorities typically have lists of all the schools or health centers in a given area. These lists can be useful when selecting a sample for a school-based or health-center-level program.

*School registers.* An experiment in Kenya on subsidies for long-lasting, insecticide-treated bed nets used school registers to create a list of households with children. This made sense for two reasons. First, primary schooling is free in Kenya, and nearly all the children are enrolled in school. Second, children are more vulnerable to malaria, so the subsidy program would target households with children, and a list constructed from the registers was as good as a census for identifying those households.<sup>26</sup> Had enrollment and attendance been low, the resulting list would have missed vulnerable households with out-of-school children.

26. This study by Jessica Cohen and Pascaline Dupas is summarized as Evaluation 6 in the appendix.



*Resource appraisals* Many programs target resources to those in need. Making the list of eligible people therefore requires us to assess who is “needy” according to the program definition. A number of ways have been developed to appraise the needy. These approaches may be used by program staff to create a list of people who are eligible for the program from which some will be randomized to receive it. They may also be used by evaluators to predict who is likely to receive the program in treatment and comparison communities. This latter approach is required if the program will assess need only in treatment communities but the evaluator needs to know who is potentially eligible in both treatment and comparison communities.

*Census with basic needs assessment.* If the program is going to target households lacking access to a given resource, such as households without a latrine at home or without a bank account, information about access to this specific resource can be measured through a census in which enumerators go from house to house and collect information on the assets of every household. A census is also useful if the program will define need based on an asset index (how many of a long list of assets a household has).

*Community-driven participatory resource appraisals.* Sometimes when eligibility is determined by outsiders based on a census as described above, the resulting list of eligible households may differ from the set of households the local community perceives to be most needy. Participatory methods capture the perceptions of the community about need. For example, a community may be asked to put forward the 10 poorest households, or the school committee or teachers may be asked to rank students by need.

*Revealed need.* The program inception can be delayed to reveal need. Imagine a program that targets scholarships to children who would not otherwise be able to attend secondary school. The school year starts in January. Instead of announcing the program and disbursing scholarships before the start of the school year, the program could be announced after the start of the school year, when all the children who would attend without help have already registered and are in school. The program can then target the qualified children who are still out of school for lack of fees.

*Hybrids.* Multiple approaches can be combined. For example, the results of the community participatory appraisal can be verified by a survey of assets, or the ranking by teachers can be checked against revealed need.

*Random sampling to create a representative list of eligible units* We may take a random sample of the population to create the pool from which we randomly allocate access to the program (the difference between random sampling and random allocation is discussed in Module 2.3). This makes sense for two reasons. First, there may be political constraints. We may have to achieve a balance of political groupings. If every political grouping (administrative district, ethnic grouping) must have a fair share, randomly sampling areas of the country to be part of the study can allow us to achieve this balance. Second, we may want to increase the external validity of the findings, so we would want the units in our evaluation to be as representative as possible of the population. (See Module 9.2.)

An education program in India did exactly this. An evaluation of a teacher performance pay program was conducted in Andhra Pradesh, which has three culturally distinct regions and 23 districts. Each district consists of five divisions, and each division has 10 to 15 subdivisions, each with about 25 villages and 40–60 government primary schools. The program was randomized at the school level. The pool of schools was created as follows: In each sociocultural region, five districts were sampled. In each district, a division was randomly selected, and in each division 10 subdivisions were randomly selected. In each of the 50 subdivisions, 10 primary schools were randomly selected using a probability proportional to school enrollment. That means that the 500 schools in the study were representative of the schooling conditions faced by a typical child in rural Andhra Pradesh.<sup>27</sup>

#### Number of randomization cells

In a simple evaluation that measures the impact of a program, we will randomize our list of eligible units into two *randomization cells*: treatment and comparison. In more complicated evaluations, we may have to divide our units into more than two cells. The number of cells will depend on the number of different treatments and the randomization strategy.

If we want to ask more complicated questions, such as which version of a program works better than another or how long the program

27. When we randomly select the sample for our study from a wider population, the whole of that wider population that is not treated is in effect part of the comparison group, even if we do not measure their outcomes. If we wanted to measure impacts more precisely later, we could go back and collect data or look at administrative outcomes from those areas that were (randomly) not picked for the study. This study by Karthik Muralidharan and Venkatesh Sundararaman is summarized as Evaluation 14 in the appendix.

has to persist before it has the effect we want to see, we have to have more than one treatment or treatment arm. We provide a number of examples of how multiple treatments can be used to answer complex questions in Module 4.6.

The research design can dictate the number of randomization cells. In the phase-in design, the number of time periods for the phase-in will determine the number of cells. For example, if the treatment is phased in over three periods, we will need three cells, one for each phase.

#### Allocation fractions

The *allocation fractions* are the proportions of the eligible units that we will assign to each group. The simplest allocation fraction is 50 percent, with 50 percent allocated to the treatment group and 50 percent to the comparison group. In most cases, dividing the sample equally between treatment and comparison maximizes statistical power. (See Module 6.4 for more details.)

#### Randomization device

We will need a randomization device. This can be mechanical (a coin, die, or ball machine), a random number table, or a computer program with a random number generator.

*Mechanical devices* We can use a mechanical device for *simple random assignment*. Examples include flipping coins, shuffling cards, rolling dice, using roulettes, picking the shortest straw, and picking names out of hats. Participants can be asked to pick a straw or to pick a ball from a transparent bowl. What should we consider before using mechanical devices? Four things: that they are ubiquitous and well accepted, they lend themselves to public randomization and transparency, they are limited to small sampling frames, and they can break down.

*Ubiquitous and well accepted.* A mechanical device has the advantage that it is perceived as fair, it can be used publicly, and it is universally familiar, used in raffles, lotteries, and other games of chance. For example, this type of device was used for the US 1969 military draft lottery for the Vietnam War.

*Lend themselves to public randomization and transparency.* Mechanical devices can facilitate public randomization and transparency. We can have the randomization in a public ceremony and can involve the participants and authorities in the randomization.

*Limited to small sampling frames.* For all their advantages, mechanical devices can be slow and tedious to use and are difficult to use with larger samples. Imagine picking names out of containers for 10,000 participants!

*Subject to mechanical failures.* Some mechanical devices will fail. For example, cards will stick together in a container, or people will grab two cards. The names put in first might never bubble to the top and never be drawn, resulting in non-equal probabilities of being drawn. These problems are real, and resolving them can undo the advantages of mechanical devices, especially the perception of fairness. For example, in the 1969 military draft lottery for the Vietnam War, the container was not well shaken, so the dates of birth put in last—those in October, November, and December—were drawn first.<sup>28</sup>

Avoid putting a large number of objects in one big sack because they may not shake and mix as well. Avoid cards that may stick together; use balls instead. Make sure that whatever you use is smooth and that someone putting a hand in the container cannot by touch know what number he is picking. Try to make the choice in a number of steps. For instance, instead of using 1,000 balls marked 000 to 999, use three containers with 10 balls apiece marked 0 to 9 and have someone choose the number by picking a ball from each urn. People may be less likely to feel cheated if the person making the selection has to draw more than once. Some mechanical devices may also be objectionable to some people because of their long association with gambling. Above all, though, keep it simple. A complicated and hard-to-understand public process will not help transparency.

*Published random number tables* A *random number table* is a list of unique numbers that are randomly ordered. Random number tables have been published from as early as 1927. For example, the RAND Corporation generated a large table by attaching a roulette wheel to a computer. The RAND table, published as the book “A Million Random Digits with 100,000 Normal Deviates,” was used for experimental design. It is good practice when using random number tables not to start at the beginning of the book or table but to randomly pick where to start from. The use of random number tables has largely been supplanted by the use of random number generators on computers.

28. T. D. Cook and D. T. Campbell, *Quasi-Experimentation: Design and Analysis for Field Settings* (Chicago: Rand McNally, 1979).

*Computerized random number generators* Random number generators are available using web-based tools, spreadsheet software, or statistical packages. For example, both Microsoft Excel and the Google Docs spreadsheet have a randomization function: `=rand()`.

#### Existing data on eligible units

We do not necessarily need to know anything about our units at the time we randomize because randomization will yield balanced treatment and comparison groups if we have a very large sample. But with smaller samples we may want to *stratify* or match units before randomizing to achieve balance on some key variables. This also helps us achieve higher statistical power and precision. Data on some key characteristics of our eligible units will also help us check whether there is indeed balance across our randomized study groups. This data may come from administrative sources or from a baseline survey we have conducted ourselves.

#### **Steps in simple random assignment**

The process of doing a simple random assignment is straightforward: we take the list of eligible units as one big pool, and we randomly assign them to different cells. If we are creating two cells (one treatment and one comparison), this can be done by flipping a coin for each unit one by one and assigning the unit to the treatment group if heads or to the comparison group if tails (or vice versa). Obviously this procedure can be quite time consuming if there are many units, and it is not possible to document that we undertook it fairly, so most evaluations use computer-based randomization. Below we describe how to do this with a spreadsheet in either Excel or Google Docs, but the procedural steps are identical if done using a statistical package such as Stata. The most commonly used process involves three simple steps:

1. Order the list of eligible units randomly.
2. Allocate units from the randomly ordered list into different groups, depending on the number of randomization cells.
3. Randomly choose which of the groups will receive treatment A, treatment B, and so on, and which will be the comparison group.<sup>29</sup>

29. It is increasingly common to avoid deciding which group is treatment versus control until the very last step.

One advantage of this three-step process is that we do not know until the very end of the process which group is considered treatment and which comparison. However, this three-step process works only when all the treatment arms are of the same size. When the probability of ending up in different treatments is different (for example, 30 percent are allocated to treatment A, 30 percent to treatment B, and 40 percent to the comparison group), we usually have to accomplish our randomization slightly differently. Both of the following approaches are valid.

1. Decide the rule by which units will be allocated to different treatment arms and the comparison group (first 30 percent to treatment A, second 30 to treatment B, last 40 to the comparison group).
2. Order the list of eligible units randomly, and allocate units to the various arms following the prestated rule.

Whatever methodology is used, it is important to agree on it and document it beforehand.

#### Order the list randomly

To randomly order the list of eligible units, we assign a unique random number to each unit on the list and then sort the list in ascending or descending order. We illustrate this in Figure 4.5 by considering the 26 letters of the alphabet as our eligible units. We enter all of the units in the sample (A–Z) in column B of an Excel or Google Docs spreadsheet. Then, in column A (not shown in the table), we type `=rand()` for all 26 rows. Every letter now has its own random number next to it. We copy all of column A and “paste special-values only” in column C. We have to copy and paste values because both Excel and Google Docs assign a new random number whenever we do anything else on the worksheet. We then select columns B and C and sort them, ascending or descending, by column C. This gives us a randomly ordered list.

#### Allocate units to groups from the randomly ordered list

When the list has been randomly ordered, we can allocate units to the groups. Say we need two groups, a treatment and a comparison group, and our allocation fractions are 50-50. We can allocate in blocks, placing the top 13 entries into Group A and the bottom 13 into

Random numbers assigned (column C) to each of the units in column B		Numbers sorted in ascending order (by column C) to get a randomly ordered list		
B	C	B	C	
A	0.257540799	W	0.095374469	Group A
B	0.141977853	P	0.127155063	
C	0.377927502	B	0.141977853	
D	0.990857584	K	0.166217843	
E	0.948417439	O	0.221630819	
F	0.303441684	Q	0.257405314	
G	0.911827709	A	0.257540799	
H	0.447802267	J	0.280958121	
I	0.287941699	I	0.287941699	
J	0.280958121	Y	0.299960876	
K	0.166217843	F	0.303441684	
L	0.871365641	C	0.377927502	
M	0.551764078	U	0.421965911	
N	0.728706001	H	0.447802267	Group B
O	0.221630819	M	0.551764078	
P	0.127155063	R	0.564626023	
Q	0.257405314	N	0.728706001	
R	0.564626023	S	0.754678177	
S	0.754678177	V	0.87089069	
T	0.907811761	L	0.871365641	
U	0.421965911	T	0.907811761	
V	0.87089069	G	0.911827709	
W	0.095374469	E	0.948417439	
X	0.987811391	Z	0.953142552	
Y	0.299960876	X	0.987811391	
Z	0.953142552	D	0.990857584	

**FIGURE 4.5** Randomizing on a spreadsheet

Group B. Or we can allocate at intervals, putting all the entries in even-numbered rows in A and all the entries in odd-numbered rows in B.

Randomly determine which group is treatment and which comparison  
In the final stage we randomly decide whether A or B is the treatment group. This can be done by flipping a coin or using a random number

generator. It is important to determine the exact procedure by which we plan to randomize before starting the process.

#### Using a statistical package

Most evaluators use a statistical package to do their randomization. The steps involved are identical to those described above. There are two key advantages in using a statistical package. First, when using complex stratification (we discuss what stratification is and why it is useful below), it is much easier to use a statistical package. More important, it is easier to record and replicate the randomization process. Statistical packages allow us to set a random seed. In other words, they pick a random number once and then remember this number. Once this is set, the program can be rerun and generate exactly the same random allocation. This allows an evaluator to prove that her allocation was random and exactly what stratification led to the random allocation that was derived. Examples of Stata code used to undertake randomization are available at the website accompanying this book ([RunningRandomizedEvaluations.com](http://RunningRandomizedEvaluations.com)).

#### Randomly order the list before selecting

Why do we need to order the list randomly? Can't we just order the list alphabetically or take the list as it was given to us and allocate every other entry to Group A and every entry between these entries to Group B? We randomly order the list because we cannot be entirely confident that any listing of the participants, no matter how haphazard, is not in fact ordered by some pattern.

For example, we may be tempted to assign people to the treatment or the comparison group based on whether their identification numbers (such as social security numbers) are odd or even. After all, it is random whether a given social security number is odd or even. However, we may worry that another program will also use odd- and even-numbered social security numbers for another evaluation. In this case, our program will not be the only systematic difference between those with odd and even ID numbers. It is unlikely, but why introduce the smallest risk into our experiment?

#### Use available data to check whether the groups we created are balanced

It has become standard practice to use baseline data (if we have it) or existing administrative data on some key observable characteristics of



our eligible units to check whether, after the randomization, the groups are balanced along those characteristics. This check reports the mean of the treatment group and the mean of the comparison group for different variables and whether they are significantly different from each other (using a *t-test*).

Table 4.3 shows a balance check performed for the study *Do Teenagers Respond to HIV Risk Information?*, which was implemented in Kenya in 2004–05.<sup>30</sup>

Why are balance checks of this kind usually included in reports on randomized evaluations? After all, if we have randomized properly, the outcome of the process (balance or no balance) must be due to chance. That is, the allocation to the treatment or the comparison group was completely independent of the potential outcomes. So why report a check?

Some economists and statisticians argue that we should not do balance checks for precisely this reason, but for the most part evaluators do include them. The main reason is that balance checks can help make the case to a skeptical reader that the program was indeed randomized. Indeed, when a researcher does not have complete control over the randomization process, a balance check can alert us to problems. For example, if randomization is being done by pulling names out of a bowl and all the names being chosen start with W, X, Y, or Z, we may worry that the bowl was not shaken properly before the drawing took place. Or maybe agricultural extension workers are asked to set up a demonstration plot in a random field, and the randomization takes place on site. If we find that the treatment fields are much closer to the center of the village than the comparison fields, we may worry that the extension workers did not follow the randomization protocol correctly and chose the most convenient fields to study. We may have to start the project again. A better solution to this problem is to use a more foolproof randomization strategy in the first place, as we describe in the next module.

Lack of balance can occur by chance even when randomization was carried out correctly. Randomization creates balance only on average, with the chance of achieving balance in any specific case increasing as the sample size increases. Often, however, we have only modest

30. Pascaline Dupas, “Do Teenagers Respond to HIV Risk Information? Evidence from a Field Experiment in Kenya,” *American Economic Journal: Applied Economics* 3 (2011): 1–34. This study is also summarized as Evaluation 4 in the appendix.

**TABLE 4.3** The balance check from an experiment on the relative risk of HIV and a teacher training experiment

School characteristics at baseline	Relative risk information			Teacher training on HIV/AIDS curriculum		
	Comparison group (C) (1)	Treatment group (T) (2)	Difference (T – C) (3)	Comparison group (C) (4)	Treatment group (T) (5)	Difference (T – C) (6)
Class size	38.2 (15.9)	34.4 (17.4)	–3.8 (1.540)**	37.4 (16.9)	37.3 (15.7)	–0.06 (1.281)
Pupils’ sex ratio (girls/boys)	1.07 (0.489)	1.12 (0.668)	0.049 (0.072)	1.06 (0.476)	1.10 (0.586)	0.040 (0.059)
Teacher/pupil ratio	0.026 (0.026)	0.026 (0.022)	0.000 (0.003)	0.025 (0.021)	0.027 (0.028)	0.003 (0.003)
Teachers’ sex ratio (females/males)	1.033 (0.914)	0.921 (0.777)	–0.112 (0.119)	1.003 (0.92)	1.014 (0.852)	0.011 (0.099)
Exam results (2003)	251.0 (29.0)	249.4 (27.4)	–1.6 (3.9)	252.2 (28.6)	249.0 (28.5)	–3.2 (3.2)

Source: Pascaline Dupas, “Do Teenagers Respond to HIV Risk Information? Evidence from a Field Experiment in Kenya,” *American Economic Journal: Applied Economics* 3 (2011): 1–34, table on 16. Reprinted by permission of the author and the American Economic Association.

Notes: The table shows the first five rows of Panel A of a multipanel table created for the above study. Standard errors are given in parentheses. \*\* indicates significant at the 5 percent level.

samples, particularly when we randomize at the group level (village, school, or district). There is a risk that the randomization can return unbalanced groups. Indeed, it is likely that if we look at a large number of different variables, we will find some significant differences in at least one variable. On average, one out of ten variables we compare across cells will be unbalanced at the 90 percent confidence level, and one out of twenty variables will be unbalanced at the 95 percent confidence level. We should not be concerned if we see this level of imbalance. In the example above, four different variables are compared between treatment and comparison groups for two different treatment groups. There is a significant difference for one variable in one of the comparisons. This is not a major problem, because it is close to what we would expect by chance and the variable, class size, is not a major factor in our theory of change.

Whether to avoid rerandomization if we do not achieve reasonable balance

What do we do if we find that, by chance, the treatment and comparison groups have very different characteristics? What if we are evaluating a business training program and our treatment group has much larger businesses than the comparison group at baseline? This may undermine our ability to draw clear conclusions from the results. If we find treatment group businesses grow faster than those in the comparison group, will we be certain that the difference is due to the program rather than to the tendency for better-capitalized larger businesses to grow faster than smaller ones? We will be able to check whether large businesses grow faster than small ones within the comparison group, or control for size of business in our analysis (see Module 8.2), but this difference at baseline is going to complicate the interpretation of our findings. The best approach is to avoid being in this situation by undertaking stratified random sampling as described in the next module.

If we do not stratify and randomization returns groups that are unbalanced along multiple dimensions of direct relevance, a common fix is to rerandomize, to try again. It used to be common practice to achieve balance by randomizing many times, selecting those random draws that achieved balance, and then randomly choosing one of these allocations for the experiment. But it is no longer considered good practice. The problem is that when we use this approach, not every combination of allocations is equally probable.

For example, if our sample happened to include one very rich person (a Bill Gates), any allocation that achieved balance would have to pair him with a large proportion of the poorest people in our sample. The other group would have to consist of the middle-income people in our sample. Whether Bill Gates and the very poor were assigned to treatment or comparison groups would be random, but certain people would always have to be in the same group to achieve balance.

The statistical rules we use to decide whether a result is significant assumes that every combination of people in the treatment and comparison groups is equally probable. But when we rerandomize, this assumption no longer holds because we reject certain combinations as unbalanced. If we deviate from this equally probable rule (as we do with stratification), we should usually account for it when we do our analysis. With stratification we know exactly what pairings we imposed:

they are the strata we constructed, and thus it is easy to control for these constraints in the analysis. With rerandomization, we can similarly control for the variables we used to balance check. However, we may not fully appreciate the precise restrictions we placed on pairings by requiring balance. In the above example, we may not realize that we forced Bill Gates and all the poorest people in the sample to be in one stratum, so we don't exactly replicate the constraints we placed on the data. It should be stressed that the practical implications of rerandomizing are unlikely to be great. However, our advice is to avoid using this technique, at least until there is more agreement in the literature about its pros and cons.

The final option is to abandon the evaluation. This is extremely costly because by the time we randomize we have usually developed a close partnership with the implementer, collected baseline data, and raised funding for the evaluation. However, it is also costly (in time and money) to continue an evaluation whose findings will be difficult to interpret.

The econometrics of stratification and rerandomization are complex, but our advice about what to do is simple. Stratified random sampling on a few key variables is usually a good idea so that we avoid the situation of having large differences between treatment and comparison groups on important variables and we can avoid rerandomization. If there are some differences between treatment and comparison groups on variables that are not the main outcomes variables of interest, it is normal and should not be considered a threat. There is no easy answer if you find large differences between the treatment and comparison groups on the baseline value of the main outcome variable, so avoid getting into that position by stratifying.

---

## **MODULE 4.5 Stratified and Pairwise Randomization**

*Stratified random assignment provides a way to ensure that our treatment and comparison groups are balanced on key variables. This module explains when stratified randomization is most useful and how to decide what variables to stratify on. It also covers pairwise randomization, which is a special form of stratified randomization.*

### **Steps in stratified random assignment**

In stratified random assignment, we first divide the pool of eligible units into strata and then within each stratum follow the procedure

for simple random assignment. The steps to perform stratified random assignment are as follows:

1. Divide the pool of eligible units into sublists (strata) based on chosen characteristics.
2. Do simple random assignment for each sublist (stratum) by
  - a. ordering the sublist randomly and
  - b. allocating units to randomization cells from the randomly ordered sublist.
3. As a final step, randomly pick which cell is treatment and which is comparison.

Imagine that we have a pool of 200 farmers, 80 men and 120 women, with half from a high-rainfall region and half from a low-rainfall region. We can first divide them by region, then by gender, and end up with four categories (Figure 4.6). We then randomly assign half of the people from each group to cell A and half to cell B. Finally, we randomize whether either cell A or cell B is the treatment group.

Thus the treatment and comparison groups are balanced by rainfall and by gender, with each cell containing 20 percent high-rainfall men, 20 percent low-rainfall men, 30 percent high-rainfall women, and 30 percent low-rainfall women. We can think of this stratification as giving chance a helping hand, because it ensures that the treatment and comparison groups coming out of the randomization are fully comparable along the stratification variables.

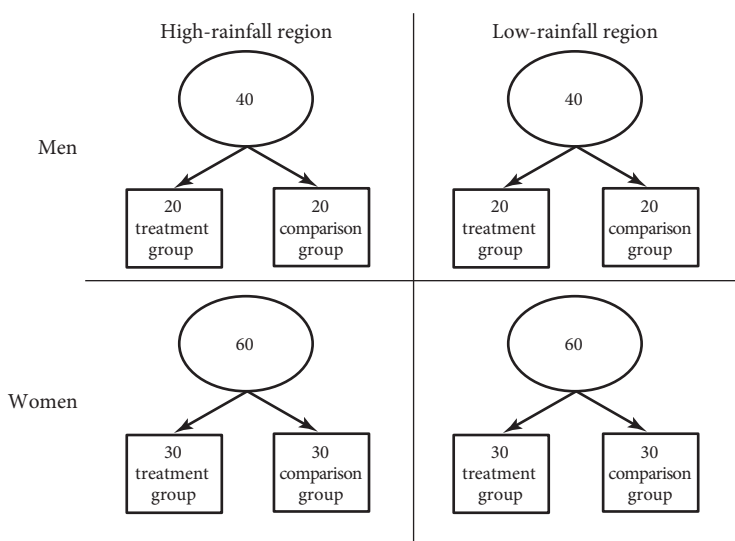
### ***When to stratify***

We stratify to achieve balance, to increase statistical power, and to facilitate analysis by subgroups. A nontechnical reason for stratification is to comply with implementer or political constraints.

*When we want to achieve balance.* The smaller the sample size, the higher the chance that simple randomization may return unbalanced groups, so the greater the need for stratification.

*When we want to increase statistical power.* As we discuss in Module 6.4, stratifying on variables that are strong predictors of the outcome can increase our statistical power.

*When we want to analyze the impact by subgroup.* We should also stratify when we want to learn how the intervention affects subgroups, such as ethnic minorities or gender. Imagine that people in our target



**FIGURE 4.6** Stratified random assignment of 200 farmers

population belong to two ethnicities, K and R, with 80 percent of the people K. Outcomes have traditionally varied by ethnicity. With simple randomization we could, in the extreme, end up with no R people (or very few) in the treatment group. Stratifying by ethnicity would ensure that both ethnicities are well represented in both groups, which would facilitate analysis by subgroup.

*When we need balance for political or logistical feasibility.* Sometimes, for political or logistical reasons, the implementers want to achieve a particular distribution of treatment groups. For example, half the treatment groups need to be in the north of the country and half in the south, or there must be exactly equal numbers of men and women. Stratification (by region, gender, or ethnicity) is a way to ensure that these constraints are met.

#### **Which stratification variables we should use**

*Variables that are discrete.* Stratification involves placing units into different buckets and then randomizing within those buckets. It is therefore not possible to stratify on a continuous variable, like income or test scores, because there might be only one person in a bucket. For example, only one child might have a test score of 67, and perhaps no child has a test score of 68. We cannot then randomize within the

bucket. Instead we put people in buckets if they have similar values for the continuous variable: for example, we group children into high-test-score or low-test-score strata. We can create even more precise test score strata by dividing children into more groups by test scores: for example, the top 10 percent of children by test score, the second 10 percent, and so on. When we say we stratify on test scores, we mean that we stratify on ranges of test scores.

*Variables that are highly correlated with the outcomes of interest.* We want balance because it simplifies the interpretation of our findings, increases their statistical power, and reduces the chance that differences in treatment and comparison groups could be driven by trends in confounding variables. Confounding variables are highly correlated with both the final outcome and participation. If available, the baseline value of the outcome of interest is a particularly important stratification variable. For example, if the outcome of interest is student test scores, a student's initial baseline score and her score at the end of the program will be correlated, so it is helpful to stratify by baseline test scores.

As we discuss below, there are usually practical constraints on the number of variables on which we can stratify. In choosing between them we should prioritize those that are most correlated with the outcomes of interest (including, potentially, the baseline value of the outcome of interest).

*Variables on which we will do subgroup analysis.* If, for example, we are going to be analyzing subgroups by gender, we will want balance on gender. We should stratify by gender to maximize power.

### ***How many variables to stratify on***

We want to stratify on variables that will explain a large share of the variation in the outcomes that interest us. But there may be many different variables that help explain the outcome. Suppose that we are measuring the impact of an education program on test scores. Age, mother's education, father's education, and baseline test score are all likely to be correlated with test scores. Do we stratify on all of them? Which do we choose? There are three considerations: stratum size, practicability, and loss of power.

*Stratum size: Allotting as many units to each stratum as there are randomization cells.* If we try to stratify on too many variables, we may find that we have strata with only one unit in them. We can't then

split the people in that stratum randomly into both treatment and comparison.

Ideally, the number of units per stratum is a multiple of the number of randomization cells. For example, with two treatment cells and one comparison, it is easier to do stratified randomization if each stratum has 9 units than if each stratum has 10. Sometimes it is impossible to avoid having strata that are not multiples of the number of cells. The best approach in this case is to randomly allocate the “leftovers” to one of the treatment arms. Sometimes with this approach we will not get exactly the same number of units in each treatment arm, which will slightly reduce our power (see Chapter 6).<sup>31</sup>

*Practicability: Trading off between balance on different variables.* The greater the number of stratification variables, the harder it may be to achieve balance on all of them. Especially when we have continuous variables, we face a trade-off between achieving tight balance on one variable or achieving a less tight balance on many variables. This is particularly true with continuous variables like test scores. Imagine that we are testing a remedial education program and have the option of stratifying on baseline test scores and ethnicity. We have 200 students in our sample. If we stratify on only test scores we create 50 strata, each of which includes only four children, two of whom would be allocated to treatment and two to control and all of whom would have very similar test scores. If we also wanted to stratify on ethnicity and there were five different ethnicities, we could not do that with our existing test score strata because there would not be two children of each ethnicity in the top test score stratum. In order to stratify on other variables, we have to stratify less tightly on test scores. The easiest way to do this is to first stratify on ethnicity and then divide each ethnicity into those with high test scores and those with low test scores. We will achieve better balance on ethnicity but less balance on test scores. We have to decide which is more important to us. If past test scores are the best predictors of future test scores, we may decide to stratify only on test scores.

*Statistical power: Trading off between low variance from more subgroups and loss of degrees of freedom from adding more restrictions.* One

31. Another strategy is to put all the “leftover” units into a stratum of their own. This approach ensures an equal number of units in every research arm, but it is controversial among some evaluators.



problem with having many strata is a potential loss of degrees of freedom in the final analysis. There is some discussion in the literature about whether we have to include dummy variables in our analysis for all our different strata (see Module 8.2). The emerging consensus is that we may not have to, but usually it is helpful to do so. If we have lots of strata and decide to control for them, there is a theoretical risk that including all these dummies could reduce our power (for more detailed discussion on covariates and the precision of our estimate of impact, see Module 8.2). But usually this is not an issue because the need to ensure that we have at least as many units in our strata as we have randomization cells severely limits the number of variables we can stratify on.

### ***The possibility of using stratified random assignment with a public lottery***

If we are performing a public randomization, stratifying is still possible. For example, if we want the program allocation to be stratified by gender, we will want men and women to pick their randomized assignment from two separate containers, each containing half treatment and half comparison markers. If we want the program to be stratified by gender and poverty status, we will want four containers: one for poor women, one for nonpoor women, one for poor men, and one for nonpoor men.

### ***Paired random assignment***

In paired random assignment, two units are matched on a list of important characteristics, and then one of them is randomly assigned to treatment and the other to comparison. If there are three randomization cells (two treatments and one comparison), the units are placed into triplets, and within each triplet we randomly assign one unit to each of the three randomization cells. Paired random assignment is stratification with strata as small as possible. Pairing may take place on one continuous variable: in the education example above, we would create 100 strata with two children in each, starting with the two children with the highest test scores and working down. Alternatively, pairing can be done by creating strata using many different variables (gender, age, high test score or low test score, north or south) until there are only as many units in each stratum as there are randomization cells.

The reasons for pairing are the same as for stratification: to achieve balance and to increase statistical power. The impact on power is greatest when there is high variability in the outcome of interest and when we have small samples. For example, pairing would be important if we were randomizing across few clusters, such as only 10 districts.

Trade-offs on the number of variables to match

As in stratification, there are trade-offs to make on the number and types of variables to pair. Only here the trade-offs are even more pronounced. The loss of degree of freedom is even stronger because pairing is an extreme form of stratification.

Attrition

Like other strategies, attrition is a threat when pairing. In paired matching, for example, if we lose one of the units in the pair (say, because the participant moves out of the area) and we include a dummy for the stratum, essentially we have to drop the other unit in the pair from the analysis. That is because the remaining unit does not have a comparison. Some evaluators have mistakenly seen this as an advantage of pairing: they suggest that if one person drops out of the study, we can drop their pair and not worry about attrition. But in fact, if we drop the pair we have just introduced even more attrition bias. This is not a good approach (and can introduce bias) because we are dropping units in response to their actions and behaviors. As we will discuss in more detail in Module 8.2, we have to stick to our initial randomization; changing what we do based on any action after this point undoes our randomization.<sup>32</sup>

Our suggestion is that if there is a risk of attrition (for example, if the randomization and pairing are at the individual level), use strata that have at least four units rather than pairwise randomization (strata with two units). If we are randomizing at the group level and there is no risk of attrition at the group level, pairwise randomization is a reasonable strategy. For example, if we randomize at the village level and

32. If we do use pairwise randomization and have attrition, we could not include dummies for every stratum and thus not drop the pair of the unit that is lost to attrition. Not including dummies for strata can be controversial (although it probably should not be) and usually reduces our power. It would be possible to set out a complicated set of rules in advance about what dummies would be included if there was attrition. However, our advice is to keep the design simple and uncontroversial.

are confident that even if we can't find specific individuals in a village we will be able to find some people to interview in that village, we don't have attrition at the level of the randomization unit.

### ***Some best practices in randomization***

Bruhn and McKenzie conducted a review of how researchers are randomizing in practice and undertook some simulations comparing how different randomization strategies performed in virtual experiments.<sup>33</sup> From their analysis they came up with recommendations on randomization procedures, including the following:

1. *Improve reporting of the random assignment method used.* Evaluators should at least include the randomization method used; the variables used to achieve balance; how many strata were used, if any; and, if rerandomizing, what the balance criteria were.
2. *Improve reporting of the randomization procedures used in practice.* How was the randomization carried out in practice? Who did it? Using what device? Was it public or private? The CONSORT guidelines, available at <http://www.consort-statement.org>, provide a structured way to report how randomization was carried out.
3. *Avoid rerandomization as a way to achieve balance.* We discussed the issues associated with rerandomization earlier in this module.
4. *Stratify on baseline outcomes, geographic location, and variables for subgroup analysis.* The gains from stratification are greatest for variables that are strongly correlated with the outcome of interest. Often many of the variables on which we want to achieve balance are correlated with geographic indicators, such as the district, and with the baseline value of our main outcome variable. Stratifying on these usually means that we achieve reasonable balance on many other variables as well.
5. *Consider statistical power when stratifying.* The more strata we have, the lower the degrees of freedom we will have when esti-

33. Miriam Bruhn and David McKenzie, "In the Pursuit of Balance: Randomization in Practice in Development Field Experiments," *American Economic Journal: Applied Economics* 1 (2009): 200–232.

mating our outcomes. Stratifying on variables that are not highly correlated with the outcome can hurt power (see Module 6.2).

We would add two more:

6. *Where possible, ensure that the randomization procedure can be replicated.* To fully document our randomization, it is useful to do it on a computer using a randomly generated seed that can be saved so that it is possible to replicate exactly how the randomization proceeded and how it was carried out.
7. *Potential attrition when stratifying.* If there is a risk of attrition at the level at which we randomize, it is good practice to include at least twice the number of units in a stratum as there are cells over which we are randomizing. In other words, with two cells (treatment and comparison) we should make sure that all strata have at least four units in them.

---

#### MODULE 4.6 A Catalog of Designs

*In this module we go over many examples of randomization in the field. These examples bring together the considerations from the four preceding modules. Each example gives the context of the study and the details of the randomization: how the opportunity to randomize came about, what was randomized and at what level, and what strategy (simple, stratified, or pairwise) was used to perform the randomized assignment.*

##### ***Treatment lottery among the eligible: The Extra Teacher Program in Kenya***

###### **Context**

In 2003 Kenya abolished primary school fees. Enrollment in primary schools rose 30 percent, from 5.9 million to 7.6 million, between 2002 and 2005. With this influx of children, class sizes in the first grade exploded. In Kenya's Western Province, for example, the average class size in first grade in 2005 was 83, and 28 percent of the classes had more than 100 pupils. The reform brought in many very poor children, and the result was not only larger classes but also classes with wide-ranging levels of preparedness.

## Interventions

*Reducing class size.* To relieve overcrowding, ICS Africa implemented the Extra Teacher Program (ETP), which funded the hiring of extra teachers. The extra teachers were recent graduates of teachers' colleges. They were given one-year renewable contracts managed by the school committee, which had full control of the contracts, including hiring, remuneration, and termination.

*Tracking students by prior achievement.* To deal with the wide range of student preparedness, a tracking component was included. Students were assigned to one of two sections based on their scores on exams given at the beginning of the year. Students with below-average test scores were assigned to one section, and those with above-average test scores were assigned to another section. It was unclear whether tracking would help or harm students. If students learn from their peers, siphoning off the high-achieving peers into the upper section would hurt students in the lower section and benefit students in the upper section. But if reducing the range of preparedness in a class helps teachers tailor the instruction to the learning levels of the students, tracking would benefit all students.

*Training school committees and formalizing their role.* To prepare the school committee for its new role of managing the contract teachers, committee members were given short, focused training on monitoring teacher performance and attendance and on soliciting input from parents. Regular meetings between committee members and district education staff were also introduced. This component was called the School-Based Management Initiative (SBM).

## Opportunity for randomized evaluation

The program was oversubscribed. ICS Africa had only enough funds to support the hiring of 120 teachers for two years. The study area included 210 rural primary schools from seven administrative districts in the Western Province.

## Level of randomization

The program could not be randomized at the student level because the intervention introduced new teachers, affecting multiple students at once. It could not be randomized by grade because the overcrowding affected mainly the first grade, where ICS Africa wanted to focus

its resources. The ETP funding was therefore randomized at the school level.

#### Questions and number of randomization cells

The effects of the ETP on learning could come from at least three channels: the reduced class size, the strong performance incentives faced by the contract teachers through the SBM, and the tracking. To disentangle the respective roles of these effects, each of the three components of the program (extra teachers, SBM, and tracking) was randomly assigned to one of three groups. The different randomization cells and the questions that could be answered by comparing different cells to each other are described in Figure 4.7.

#### Randomization

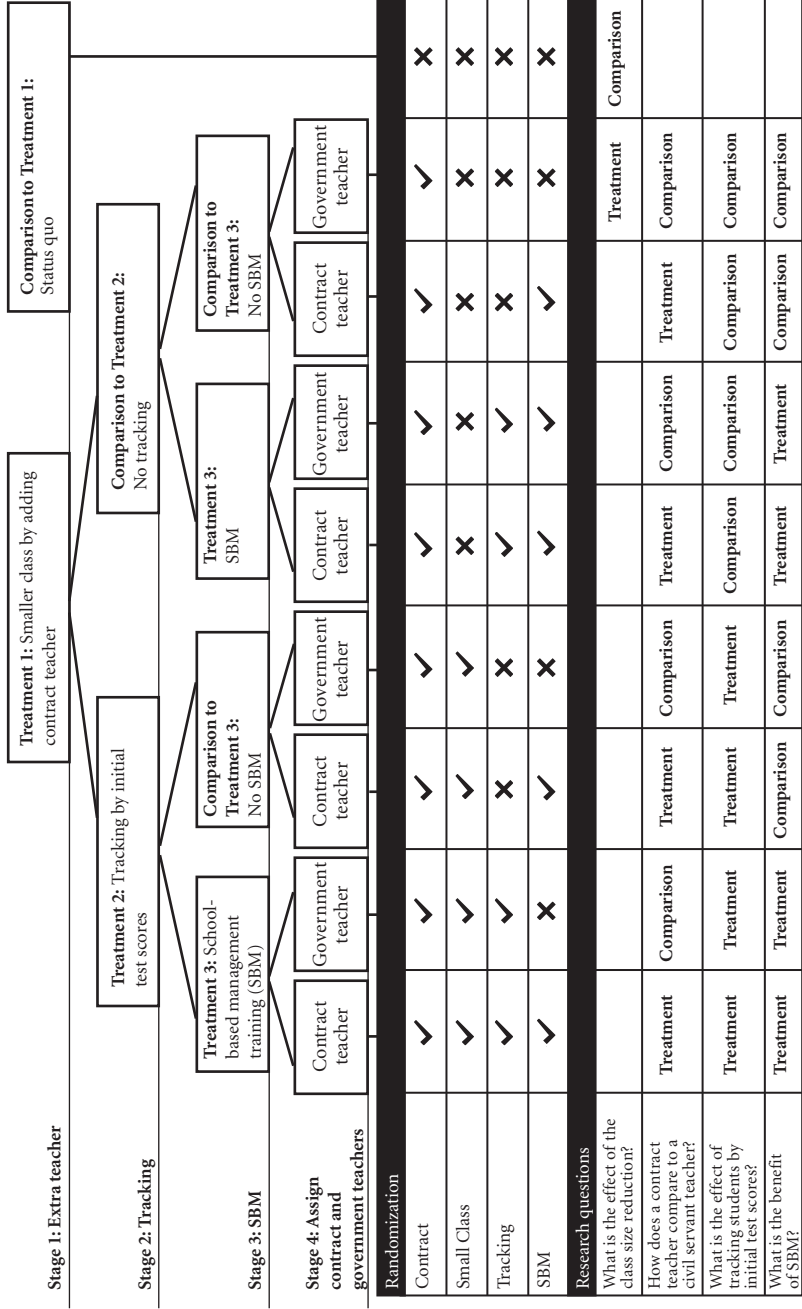
*Stratification.* Baseline data on enrollment, the pupil/teacher ratio, and the number of grade 1 sections were collected in 2004, prior to the start of the program. The schools were stratified by administrative district and by the number of grade 1 sections. There were a total of 14 strata.

*Allocation.* The schools were assigned to randomization cells in a multistage lottery using a computer.

*Stage 1: Randomizing the ETP.* There were 210 schools in 14 strata. The allocation fraction for the ETP program was two-thirds treatment (to receive funds for an extra teacher) and one-third comparison. In each stratum the schools were ordered randomly; the top two-thirds were assigned to receive funds, and the bottom third was assigned to the comparison group. In total, 140 schools received funds for an extra teacher. More schools were allocated to the treatment than to the comparison group because the treatment group was then going to be subdivided into several different randomization cells, and it was important to have sufficient samples in each to be able to compare different treatments to each other and to the comparison group.

*Stage 2: Randomizing into tracking.* Of those schools allocated to the ETP, half were allocated to the tracking treatment through the same stratified random ordering process described above.

*Stage 3: Randomizing the SBM program.* Two pools of schools were assigned to SBM or non-SBM: the 70 tracking schools and the 70 non-tracking schools. Half the schools in the tracking pool were randomly assigned to the SBM program. Similarly, half the schools in the non-tracking pool were randomly assigned to the SBM program. In other



**FIGURE 4.7** Randomization cells and research questions for the Extra Teacher Program in Kenya

words, allocation to tracking was used as a variable for stratifying allocation to the SBM program.

*Stage 4: Within-school randomization of teachers to sections.* In addition, the teachers were randomly assigned to the different sections. Once an additional section had been created and the students had been assigned to sections, either with or without tracking, the contract teachers were randomly assigned to teach one section to make sure that the regular teacher did not always take the higher-level class when classes were tracked. The other sections were taught by the regular civil-service teacher.

For further reading

Duflo, Esther, Pascaline Dupas, and Michael Kremer. 2012. "School Governance, Teacher Incentives, and Pupil-Teacher Ratios: Experimental Evidence from Kenyan Primary Schools." NBER Working Paper 17939. National Bureau of Economic Research, Cambridge, MA.

This study is summarized as Evaluation 3 in the Appendix.

### ***Treatment lottery among the marginally ineligible:***

#### ***Credit in the Philippines***

##### **Context**

First Macro Bank, a rural bank in the Philippines, provides business and consumer loans to poor clients. The bank uses a credit scoring algorithm based on the following characteristics: business acumen, personal financial resources, recourse to outside finances, personal and business stability, and demographic profile. Scores range from zero to 100. Applicants scoring below 31 are not considered creditworthy and are rejected automatically. The creditworthy applicants fall into two categories: the very creditworthy, those scoring above 59, who are automatically approved for a loan, and the marginally creditworthy, those scoring between 31 and 59.

##### **Opportunity for randomized evaluation**

The bank piloted a program to expand its services to a new client population: the marginally creditworthy. This program expanded credit to poor clients who might not otherwise receive credit. It also allowed the bank to gather data on these clients, which could be used to improve its credit-scoring model, risk management, and profitability.



## Creating a sampling frame

*Initial screening.* Loan officers screened applicants for eligibility. To qualify, an applicant had to (1) be aged between 18 and 60 years, (2) have been in business for at least one year, (3) have been resident at his present address for at least a year if an owner and for at least three years if a renter, and (4) have a daily income of at least 750 pesos. Some 2,158 of the applicants passed the initial screening.

*Credit scoring.* Business and household information on these 2,158 applicants was entered into the credit-scoring software. Of these, 166 applicants scored between zero and 30 and were automatically rejected, 391 scored between 60 and 100 and were automatically accepted, and 1,601 scored between 31 and 59. These 1,601 were the sampling frame and were allocated randomly (see Figure 4.4).

## Randomization

*Stratifying by credit score and using different allocation fractions in the strata.* The random assignment still took into account the credit scores. Among the 256 applicants with scores between 31 and 45, the probability of receiving a loan was 60 percent, but among the 1,345 applicants with scores between 46 and 59, the probability was 85 percent.

*Allocation.* Altogether, of the 1,601 applicants in this range, 1,272 were assigned a loan approval, and 329 were assigned a rejection. To reduce the chance that clients or loan officers would change the applications to improve the applicants' chances of getting a loan, neither group was informed of the algorithm or of its random component.

*Verification of applicant information and final decision.* The credit-score-based decisions were conditioned on verification of applicant information. Verification included a visit to each applicant's home and business, meeting neighborhood officials, and checking references.

## For further reading

Karlan, Dean, and Jonathan Zinman. 2011. "Microcredit in Theory and Practice: Using Randomized Credit Scoring for Impact Evaluation," *Science* 332 (6035): 1278–1284.

This study is summarized as Evaluation 13 in the appendix.

## ***Randomized phase-in: The Primary School Deworming Project in Kenya Context***

Worms affect more than 2 billion people in the world, causing, among other symptoms, listlessness, diarrhea, abdominal pain, and anemia.

Worldwide, 400 million school-aged children are at risk of worm infection. The World Health Organization (WHO) recommends preventive school-based mass treatment in areas with high levels of worm prevalence. Treatment with deworming pills kills worms in the body. Although this reduces transmission, it does not prevent re-infection. Schools with a hookworm, whipworm, and roundworm prevalence of more than 50 percent should be mass treated every six months, and schools with a schistosomiasis prevalence of more than 30 percent should be mass treated once a year. ICS implemented a school-based mass deworming program that treated 30,000 pupils at 75 primary schools in Kenya's Western Province.

#### Opportunity for randomized evaluation

This was a pilot program; a key goal was to see whether, and how, deworming affected education. The program faced logistical and financial constraints.

The logistics were particularly complicated. The medication had to be acquired and parental informed consent to treat children obtained by having parents sign a ledger at the school. The prevalence of worms in the area had to be tested to see if a school qualified for mass treatment, and ICS public health officers and nurses seconded from the Ministry of Health had to be trained. Treatment dates had to be worked out with the schools, avoiding the two rainy seasons when the roads were hardly passable. In most cases children had to be treated twice a year because of the high infection rate in the area. Two categories of children who could not be treated had to be separated out: those whose parents had not given consent and all girls of reproductive age. (At the time it was thought that there was a risk of birth defects. It has now been found that there is no risk, and the recommendation is to treat all children.) Coverage could not be extended to all schools at once; for some of the schools, treatment had to be delayed. These logistical constraints made a phase-in design the most feasible.

#### Level of randomization

Most schools in the area met the WHO guidelines for mass treatment. The school was the natural unit of intervention for the school-based mass treatment program. Since worms are easily transmitted between children, spillovers were a major consideration. When there are no latrines, children relieve themselves in the bush around the school or

around the home, creating the potential for spillovers around schools and homes. Randomizing at the school level would capture within-school spillovers in the impact estimate. The schools were far enough from each other that at least some schools would not be influenced by spillovers. The randomization was not heavily stratified to ensure that there was variation in treatment density in different geographic areas, allowing for measurement spillover effects (Module 8.2 discusses how the analysis measured spillovers in some detail).

#### Considerations for a phase-in design

Logistical constraints meant that ICS had to phase in the program, suggesting a randomized phase-in design. But it was unclear what the gap between different phases should be. There were two considerations: how fast could the program be phased in, and what was the likely timeline of impact?

*Logistical constraints and resource availability.* There would be enough financial resources for all schools by the beginning of the fourth year. Creating three phase-in groups with 25 schools each allowed ICS to benefit from economies of scale in training teachers for the program but did not overstretch their capacity.

*Likely timeline of impact.* Worms in the body would be killed immediately, but some of the benefits would become apparent only over time (for example, if children learned more or grew faster as a result of treatment). If the time between phases was very short, it might be difficult to pick up these benefits.

#### Randomization

*Stratification.* The schools were first stratified by administrative district and then by their involvement in other nongovernmental assistance programs.

*Allocation.* The schools were then listed alphabetically, and every third school was assigned to a group. This divided the 75 schools into three groups of 25. As discussed in Module 4.1, this strategy is less ideal than pure randomization, but in this case it arguably mimicked randomization.<sup>34</sup> Next the three groups of schools were phased into the program. In the first year, Group 1 schools were treated, while Groups 2 and 3 formed the comparison. In the second year, Group 2

34. The concern here would be if someone else used alphabetical order of schools to allocate their program. There is no reason to think that that occurred in this case.

was phased in. Now Groups 1 and 2 formed the treatment group, while Group 3 remained the comparison. In the fourth year, Group 3 was also treated, which ended the evaluation period because no comparison group remained.

For further reading

Baird, Sarah, Joan Hamory Hicks, Michael Kremer, and Edward Miguel. 2011. "Worms at Work: Long-run Impacts of Child Health Gains." Working paper, Harvard University, Cambridge, MA. [http://elsa.berkeley.edu/~emiguel/pdfs/miguel\\_wormsatwork.pdf](http://elsa.berkeley.edu/~emiguel/pdfs/miguel_wormsatwork.pdf).

J-PAL Policy Bulletin. 2012. "Deworming: A Best Buy for Development." Abdul Latif Jameel Poverty Action Lab, Cambridge, MA. <http://www.povertyactionlab.org/publication/deworming-best-buy-development>. This study is summarized as Evaluation 1 in the appendix.

### ***Rotation: Remedial education in India***

#### **Context**

In 1994, Pratham, an Indian organization working in education, launched a remedial education program. Pratham hired and trained tutors and deployed them to schools. The tutors worked with children who had reached grades 3 and 4 without mastering the basic reading and math competencies taught in grades 1 and 2. The lagging children were identified by their teachers and were pulled out of their regular classes in groups of 20 and sent for remedial tutoring for half the school day.

#### **Opportunity for randomized evaluation**

In 2000 Pratham expanded their program to primary schools in the city of Vadodara in western India. The program was expanding to a new area and reaching a new population. There were not enough resources to cover all the schools at once, but the municipal authorities requested that all eligible schools receive program assistance.

#### **Level of randomization**

The municipal government and Pratham agreed that all the schools in the new area of expansion that needed tutors should receive them. This meant that the program could not be randomized at the school level. Nor could the program be randomized at the student level. Because this was a pullout program, randomizing at the student level would have meant that the teacher would identify the lagging students, and then half of those would be selected randomly. Both

logistical considerations ruled this out. Instead, tutors were randomly assigned to a specific cohort, either grade 3 or grade 4. Every school still received a tutor but for only one of the two grades.

#### Randomization

*Stratification.* The schools were stratified by language of instruction, pretest scores, and gender.

*Allocation.* The schools were randomly assigned to two groups. In the first year, Group A would receive tutors in grade 3 and Group B would get them in grade 4 (Figure 4.8). In the second year, the groups would switch: Group A would receive tutors for grade 4 and Group B would get them for grade 3. After the first year of the program, the researchers could measure the impact of having a tutor for one year. Group A schools' grade 3 with a tutor were compared to Group B schools' grade 3. The reverse was true for grade 4.

After one year, the schools switched, with the tutors moving to grade 4 in Group A schools and the tutors in Group B moving to grade 3 (Figure 4.9). Because children progress from one grade to the next, this meant that children in Group A schools who moved from grade 3 to grade 4 had a tutor for two consecutive years, while their peers in Group B who moved to grade 4 did not have a tutor at all. This allowed the researchers to measure the impact of having a tutor for two years.

#### For further reading

Banerjee, Abhijit, Shawn Cole, Esther Duflo, and Leigh Linden. 2007. "Remedying Education: Evidence from Two Randomized Experiments in India," *Quarterly Journal of Economics* 122 (3): 1235–1264.

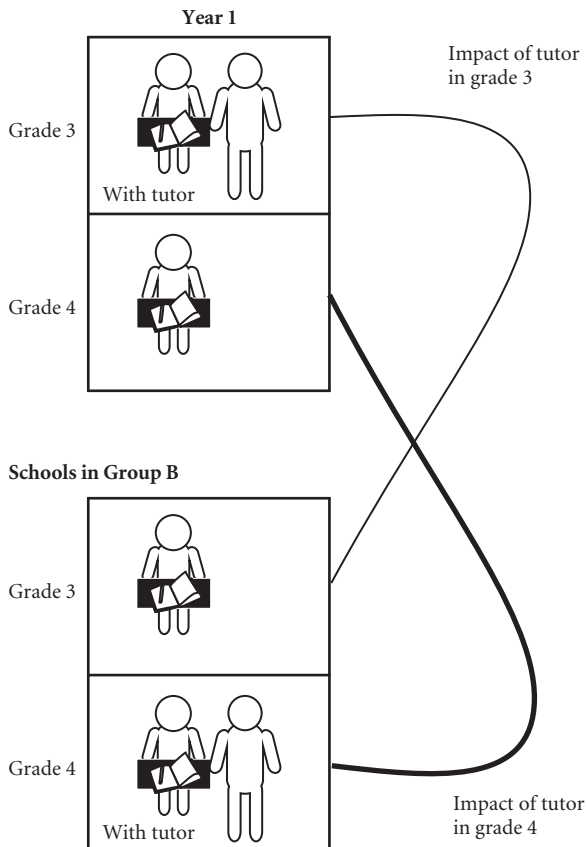
J-PAL Policy Briefcase. 2006. "Making Schools Work for Marginalized Children: Evidence from an Inexpensive and Effective Program in India." Abdul Latif Jameel Poverty Action Lab, Cambridge, MA. <http://www.povertyactionlab.org/publication/making-schools-work-marginalized-children>. This study is summarized as Evaluation 2 in the appendix.

### ***Encouragement: Retirement savings at a large American university***

#### Context

To increase savings for retirement, many employers in the United States offer to match employees' retirement contributions. Savings in these accounts also have lower tax rates than other savings accounts. But despite these incentives, many employees do not sign up for employer-matched retirement accounts. At a large American university, only 34 percent of eligible employees were enrolled. Because one

### Schools in Group A



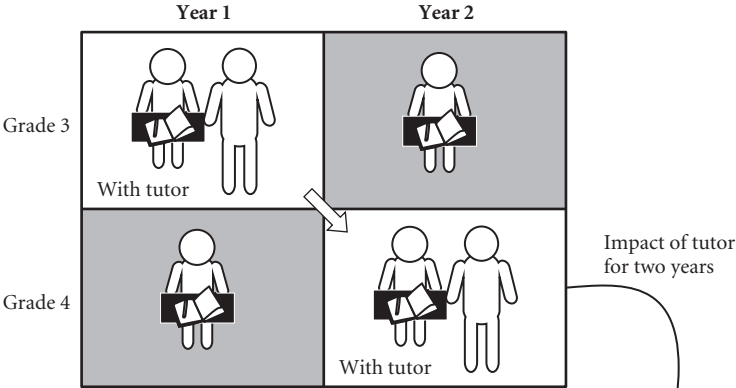
**FIGURE 4.8** Rotation design: Measuring the impact of having a tutor for one year

reason employees did not sign up may be that they are unaware of the extent of the benefits, the university holds a fair yearly to educate employees about benefits.

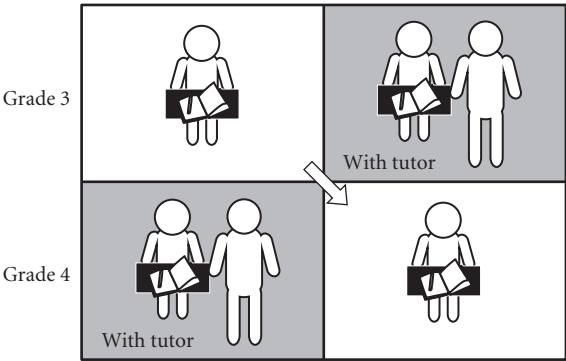
#### Opportunity for randomized evaluation

The information fair was open to all university employees, but attendance was low. This meant there was room to increase take-up by offering employees a \$20 reward for attending. This extra encouragement could be randomized.

Schools in Group A



Schools in Group B



**FIGURE 4.9** Rotation design: Measuring the impact of having a tutor for two years

Level of randomization

The fair targeted individual employees, who would decide whether to enroll for benefits and how much to contribute. But an employee's decision to attend the fair and to enroll in the account could depend on social interactions with colleagues. These spillovers could be both informational and behavioral. Employees could remind each other to attend the fair. The fair was held at two hotels some distance from the school, and people tended to go to the fair in groups by department. Someone was more likely to attend if there was a group from the department going. People who attended the fair could share informa-

tion from the fair with people who did not attend. People could also mimic the investment decisions of their colleagues. The investment decisions themselves may have depended on social norms or beliefs about social norms. People may have wanted to conform to the savings norms they observed in their department.

Randomizing at the department level would capture the potential within-department spillovers, but savings could be measured at the individual level. Because the evaluators were interested in the social dynamics of saving, they randomized both at the department and the individual levels. There were 330 departments and 9,700 employees, of whom 6,200 were not enrolled in the retirement savings accounts. A week before the fair, a letter was sent to a randomly selected subset of employees to remind them about the fair and to inform them that they would receive a check for \$20 if they attended.

#### Considerations for using the encouragement design

Why use money as an encouragement? Money is a good inducement because it is unrestricted and thus likely to draw the widest range of people in the population. By way of comparison, offering a free football or T-shirt would attract only certain types of people. The sum of \$20 was likely chosen because it was too small to affect the level of savings on its own (the minimum annual contribution was \$450) but large enough to encourage attendance at the fair. There might have been a concern that a \$20 incentive would be more likely to attract those with lower salaries, but because these were the employees who were least likely to have employer-matched accounts, they were the main targets for the program.

#### Questions and number of groups

The objective was to isolate both the effects of peers on savings behavior and the effects of the fair on savings. This suggests randomization in at least two stages: one to create the experimental groups needed to isolate the social effects in the departments and the other to isolate the individual effects.

#### Randomization

*Randomizing pairwise.* The 330 departments were ranked by participation rates among staff and divided into 10 groups of 33. Within each group, the 33 departments were then ranked by size. Once ranked, the



departments were divided into 11 groups of three, each containing three consecutive departments.

*Randomizing the departments to estimate social effects.* Within each triplet, two departments were randomly allocated to the treatment. In total, 220 departments were in the treatment group.

*Randomizing individuals within departments to isolate individual effects.* There were 6,200 employees who had not enrolled scattered in these departments. Of these, 4,168 were in departments assigned to treatment. Within each of the treatment departments, half of the employees were randomly assigned to receive the encouragement. This means that there were two treatments: a direct treatment (being encouraged to attend the fair) and an indirect treatment (being in the same department as someone encouraged to attend the fair). In the end, 2,039 employees received an encouragement letter one week before the fair, 2,039 employees had a colleague who had received encouragement, and 2,043 employees were in the 110 comparison group departments.

For further reading

Duflo, Esther, and Emmanuel Saez. 2003. "The Role of Information and Social Interactions in Retirement Plan Decisions: Evidence from a Randomized Experiment." *Quarterly Journal of Economics* 118 (3): 815–842.

This study is summarized as Evaluation 11 in the appendix.

### ***Randomization as an instrumental variable: Delayed marriage incentive in Bangladesh***

#### **Context**

Bangladesh has one of the highest rates of adolescent and child marriage in the world. Although the legal age of marriage for women is 18, it is estimated that nearly 50 percent of all girls and 75 percent of rural girls are married by age 15. Women who marry early have worse education, income, and maternal and child health outcomes. But does early marriage cause these negative outcomes, or is it a symptom of another underlying problem, such as poverty? To isolate the effect of early marriage we need to have random variation in age of marriage.

We cannot randomly assign the age of marriage, just as we cannot randomize many other variables of interest to policy. But we can randomize other variables that may influence the age of marriage as long as they do not themselves affect the outcome of interest. These variables are called instrumental variables, and they are very similar to

the encouragement approach discussed above. Here the evaluators, working with Save the Children, mimicked delayed marriage by providing financial incentives to families of unmarried adolescent girls. If the incentives worked, randomized assignment of the incentives would generate the random variation in age of marriage needed to rigorously estimate the effects of early marriage.

### Intervention

Research in Bangladesh has suggested that for each year a family delays the marriage of an adolescent girl, the dowry they have to pay increases by about US\$15, creating a financial incentive for early marriage. In a random sample of villages in rural Bangladesh, Save the Children gave families with unmarried girls aged 15–17 a monthly transfer of cooking oil with a yearly value close to US\$15. The incentive was given only to families of 15- to 17-year-olds because the funding for the incentive was severely limited and this was the age group with the highest marriage rate. A larger incentive might have had a greater impact on the age of marriage, which would have been useful in assessing the impacts of changing the age of marriage. But too large an incentive might have started to have direct impacts by changing the families' income in a significant way, which would have made it impossible to isolate the impact of change in the age of marriage from that of change in income. In addition, there were not sufficient funds to increase the incentive.

The delayed marriage incentive was part of a larger pilot empowerment program. That program had two versions. The basic package consisted of the establishment by the community of safe spaces where the girls could meet, such as the veranda of a home or a room in a school. In this safe space, girls received health education, homework support, and general empowerment training. A separate variant of the program added a financial readiness component to the training curriculum.

### Opportunity for randomized evaluation

This pilot incentive program was designed to piggyback on an existing program working on the provision of food security and nutritional support to pregnant women and nursing mothers. But there were not sufficient resources to implement the adolescent girls' program everywhere the food security program operated, which created an opportunity for randomization.

## Level of randomization

There were many considerations in choosing the level of randomization. For the delayed marriage incentive, individual randomization was ruled out because of the risk that this would create confusion and resentment. Individual randomization was also not possible for the safe spaces program because the community would decide the location of appropriate safe spaces. This meant that the program had to approach and sign up potential participants at the community level.

However, there was some flexibility about what a “community” was. We tend to think of a community as a simple and clearly defined geographical unit, but in practice there are often multiple ways for people to come together in a community. Deciding that the level of randomization is the community is just the first step. This is a particular concern in densely populated areas, such as rural Bangladesh or urban slums, where one community blends into the next. Adjacent neighbors may live in different administrative units, or the next farm in a village may be in another country.

One clearly identifiable unit was the cluster of houses around a common courtyard that make up a *bari*. Because there is substantial interaction within a *bari*, it could be considered a community. But a *bari* would not have enough adolescent girls for a viable safe space program.

Another option was to use the community units that were used by the food security and nutrition program to implement their intervention (i.e., to define which households should go to which food distribution center). However, if the nutrition program’s concept of community were used, only a few communities could be covered because of resource constraints. There would not be enough different units for a good evaluation.

Another factor in the decision on randomization level was potential general equilibrium effects. If suddenly all the girls in an area married later, there could be an impact on the marriage market, particularly if men have a strong preference for marrying girls from the local area. There could be a short-run shortage of girls with a sudden shift in the age of marriage, which would even out if the delay in marriage persisted. It would be good to design the study to avoid picking this effect up, suggesting a unit of randomization that was smaller than the marriage market. There might be other general

equilibrium concerns that would suggest a level of randomization at the marriage market level. For example, if girls were more educated when they married, this might mean that their families would seek more educated husbands. If the randomization level was smaller than the marriage market, treatment girls would marry more educated men in a way that would not be possible if the program were scaled up (it is unlikely that the program would lead to an increase in the number of educated men). However, it was decided that randomizing at the marriage market level would be too expensive. Instead, details on the characteristics of the husbands would be carefully documented to track whether the program led to changes in husband selection.

In the end, the evaluation identified “natural villages,” geographic groupings that were small enough to allow a large sample size for evaluation but also represented groupings around which people on the ground would naturally organize themselves. Although most of these natural villages were not separate administrative units within the government system, they were recognized in addresses. The randomization was based on these natural villages. This approach, however, involved some risk that the lines between natural villages would not be clear enough in people’s minds and that some girls from comparison villages would end up attending the safe space program, a risk that did in the end materialize.

#### Randomization

Altogether, communities were divided into six experimental groups of roughly 77 girls each; the first two were both comparison groups because of the importance the researchers gave to finding the precise impacts of the individual components of the program. So the groups were designated as follows: (1, 2) comparison; (3) safe spaces and education support; (4) safe spaces, education, and financial readiness; (5) safe spaces, education, financial readiness, and delayed marriage incentive; and (6) delayed marriage incentive. The randomization was performed using Stata and involved stratifying by union (a geographical cluster of roughly 10 villages). Within strata, villages were ordered by size. A random number between one and six was chosen: if the number six was chosen, the first community in the stratum was allocated to the delayed marriage incentive, the next two to comparison, the next to safe space with education, and so on.

For further reading

This study by Erica Field and Rachel Glennerster is summarized as Evaluation 7 in the appendix.

***Creating variation to measure spillovers: Job counseling in France***

To measure spillovers, we need variation in exposure to spillovers, which can be achieved by having variation in the treatment density. Treatment density is the proportion of units (individuals, schools, or communities) within a geographic area that receive the treatment. The higher the treatment density, the higher the chances of spillover effects on the average untreated person. We need to have some of our comparison group that we have good reason to believe do not experience spillovers; these will be the “pure comparison” group. In our earlier discussion of the deworming evaluation under randomized phase-in, we described how evaluators can use chance variation in treatment density to measure spillovers. Here we discuss an example in which evaluators have deliberately created (random) variation in treatment density. Another example is the study of savings in an American university discussed above, in which different departments were randomized to receive different intensities of encouragement.

Randomizing the treatment density

We can directly vary the treatment density. A study in France wanted to measure the negative spillover effects of job counseling. Here the fear is that providing counseling will help the counseled secure employment at the expense of those without counseling—that is, that there will be negative spillovers. Because the goal of the policy was to reduce unemployment, it was important to measure any negative spillovers.

The program was conducted through public employment agencies. There were 235 of these in 10 administrative regions. The randomization was at the employment agency level. Each agency was considered a small and autonomous labor market. The agencies were placed into groups of five based on size and population characteristics. There were 47 such groups. Within each group, the five agencies were randomized to one of the five treatment densities: zero percent, 25 percent, 50 percent, 75 percent, or 100 percent of the unemployed applicants were assigned to the treatment group. The unemployed were randomized as they applied. In agencies with zero percent, none

were treated; in those with 25 percent, a quarter were treated, and so on. The study examined whether an untreated (not counseled) unemployed person in an agency where 75 percent of the unemployed were treated (counseled) has a *lower* chance of obtaining employment than an untreated unemployed person in an agency where only 25 percent of the unemployed were treated. In other words, were the untreated people in areas where they had to compete with a higher proportion of counseled people worse off? If yes, there were negative spillover effects.

For further reading

- Crépon, Bruno, Esther Duflo, Marc Gurgand, Roland Rathelot, and Philippe Zamora. 2012. "Do Labor Market Policies Have Displacement Effects? Evidence from a Clustered Randomized Experiment." NBER Working Paper 18597. National Bureau of Economic Research, Cambridge, MA.
- . 2011. "L'Accompagnement des Jeunes Diplômés Demandeurs d'Emploi par des Opérateurs Privés de Placement." *Dares Analyses* 94: 1–14. This study is summarized as Evaluation 8 in the appendix.