J-PAL
ABDUL LATIF JAMEEL POVERTY ACTION LAB
NORTH AMERICA

REAL-WORLD CHALLENGES TO RANDOMIZATION AND THEIR SOLUTIONS

Kenya Heard, Elisabeth O'Toole, Rohit Naimpally, Lindsey Bressler

J-PAL North America, April 2017

povertyactionlab.org/na

Purpose: This tool is primarily intended for policymakers and practitioners who have a general understanding of randomized evaluations and want to learn how to address six common challenges. Throughout, we will reference evaluations by J-PAL affiliates and their co-investigators. This document draws from *Running Randomized Evaluations: A Practical Guide* by Rachel Glennerster and Kudzai Takavarasha (henceforth referred to as GT).

Acknowledgements: We are grateful to Mary Ann Bates, Joseph Ciesielski, Laura Feeney, Amy Finkelstein, Christina Galardi, Rachel Glennerster, Sara Heller, William Pariente, and Marc Shotland for providing thoughtful feedback to improve this resource. Many thanks to Rachel Glennerster and Kudzai Takavarasha for allowing us to draw from their book. We thank the authors of the studies, Katherine Baicker, Bruno Crépon, Esther Duflo, Robert Fairlie, Marc Gurgand, Amy Finkelstein, Matthew Notowidigdo, Roland Rathelot, Jonathan Robinson, and Philippe Zamora, for allowing us to use their evaluations as real-world examples to bring concepts to life. Graham Simpson copy edited this document, and Alicia Doyon formatted the document and designed the figures.

INTRODUCTION

Randomized evaluations, also called randomized controlled trials (RCTs), have received increasing attention from practitioners, policymakers, and researchers due to their high credibility in estimating the causal impacts of programs and policies. In a randomized evaluation, a random selection of individuals from a sample pool is offered a program or service, while the remainder of the pool does not receive an offer to participate in the program or service. Random assignment ensures that, with a large enough sample size, the two groups (treatment and control) are similar on average before the start of the program. Since members of the groups do not differ systematically at the outset of the experiment, any difference that subsequently arises between the groups can be attributed to the intervention rather than to other factors.

Researchers, practitioners, and policymakers face many real-world challenges while designing and implementing randomized evaluations. Fortunately, several of these challenges can be addressed by designing a randomized evaluation that accommodates existing programs and addresses implementation challenges.

Program design challenges: Certain features of a program may present challenges to using a randomized evaluation design. This document showcases four of these program features and demonstrates how to alter the design of an evaluation to accommodate them.

- Resources exist to extend the program to everyone in the study area
- Program has strict eligibility criteria
- Program is an entitlement
- Sample size is small

Implementation challenges: There are a few challenges that may threaten a randomized evaluation when a program or policy is being implemented. This document features two implementation challenges and demonstrates how to design a randomized evaluation that mitigates threats and eliminates difficulties in the implementation phase of an evaluation.

- It is difficult for service providers to adhere to random assignment due to logistical or political reasons
- The control group finds out about the treatment, benefits from the treatment, or is harmed by the treatment

A note on statistical power

Many of the challenges herein will reference the concept of *statistical power*, which is synonymous with *power*. The power of an evaluation reflects the likelihood of detecting any meaningful changes in an outcome of interest brought about by a successful program. Power is determined by many factors including the sample size and anticipated effect of the program. This document will mention but will not detail statistical power, so for more information, please refer to J-PAL's research resource: The Danger of Underpowered Evaluations.

TABLE OF CONTENTS

PROGRAM DESIGN CHALLENGES

Challenge #1: Resources exist to extend the program to everyone in the study area

Overview

In some cases, program implementers have enough resources and may feel obligated to distribute a program to everyone in the study area. If everyone receives the treatment at once, there are not any untreated individuals to make a control group for a randomized evaluation.

CHALLENGE
Resources exist to extend the program to everyone in the study area.

IMPLICATION
If everyone is treated at once, there is no control group.

SOLUTION
Use a phase-in design and randomly assign the order in which groups or individuals receive treatment.

A Real-World Example

Researchers Robert Fairlie (University of California, Santa Cruz) and Jonathan Robinson (University of California, Santa Cruz) faced this challenge when they conducted an evaluation to determine the impact of home computers on academic achievement. Under the program, students who did not have a home computer were eligible to receive a computer. To conduct the evaluation, the researchers needed to compare the outcomes of eligible students who received a computer and eligible students who did not receive a computer. The researchers talked with school officials about the study design and concluded that it would be unfair to give some eligible students computers and withhold computers from the rest (Fairlie and Robinson 2013, 215).

Solution: Phase-in

To ensure a fair distribution of computers while preserving a control group, researchers used a phase-in design. The treatment group received computers at the beginning of the school year, and the control group received computers at the end of the school year. An endline survey was administered at the end of the school year to measure impacts of the program just before control students received their computers. By the end of the school year, every eligible student received a computer.

Advocates for the distribution of home computers might have argued that the phase-in design was not fair because control students had to wait a year to receive computers. However, it is important to remember that the effect of home computers was not known. Even though there were enough resources to treat everyone, if every student received a computer at the beginning of the school year, there would have been no way to determine if this expensive program had positive, negative, or no effects on students. Without an impact estimate, there would be no way to determine whether another intervention that aimed to improve educational outcomes would have a larger impact and be more cost-effective than a home computer program.

The phase-in design allows the control group to remain an appropriate representation of the **counterfactual** while ensuring fairness when there are resources readily available to treat everyone. This study used two phases, but it is also possible to use multiple phases as shown in Figure 1.1. Researchers might decide to use multiple phases depending on several factors that shape the context of the evaluation (e.g., the length of the intervention, the number of randomization units, the particular outcomes they desire to study). Additionally, if a randomized evaluation occurs as a program expands its capacity, there may be a need to train additional staff or troubleshoot logistical challenges that come with scale. In these cases, instead of scaling up and implementing the program all at once, the phase-in design provides extra time to scale up carefully while maintaining fidelity to the program model.

FIGURE 1.1: PHASE-IN STUDY DESIGN

Limitations

- *Anticipation of treatment may affect the behavior of the control group*. Say some parents of children in the control group were planning to purchase a home computer prior to the evaluation. However, knowing that their child would receive a home computer at the end of the year, they may have decided not to purchase a home computer. This behavioral change is important because some control group parents behaved differently than they would have if the program did not exist. Their decisions were affected by the intervention itself, which means they are no longer the best representation of the **counterfactual**. If this is the case, impact estimates may overestimate the effects of home computers. These changes in behavior, which can lead to an overestimate or an underestimate of impact, are called **anticipatory effects**.
- *Since the control group receives treatment after a fixed time frame, there is a limited time over which impact can be measured.* Researchers studying the home computer program only evaluated the program's impact on short-term outcomes (i.e., since all students received a computer and were "treated" after one year, researchers could only evaluate the one-year impact of the program on students' outcomes). Since everyone eventually

receives the program and the control group becomes a part of the treatment group at the conclusion of the evaluation, it is challenging to observe long-term impacts of the program.

An exception to this limitation occurs in a setting where individuals move in and out of the randomization unit. In this case, researchers could determine the long-term impact of the home computer program if the intervention were structured differently.

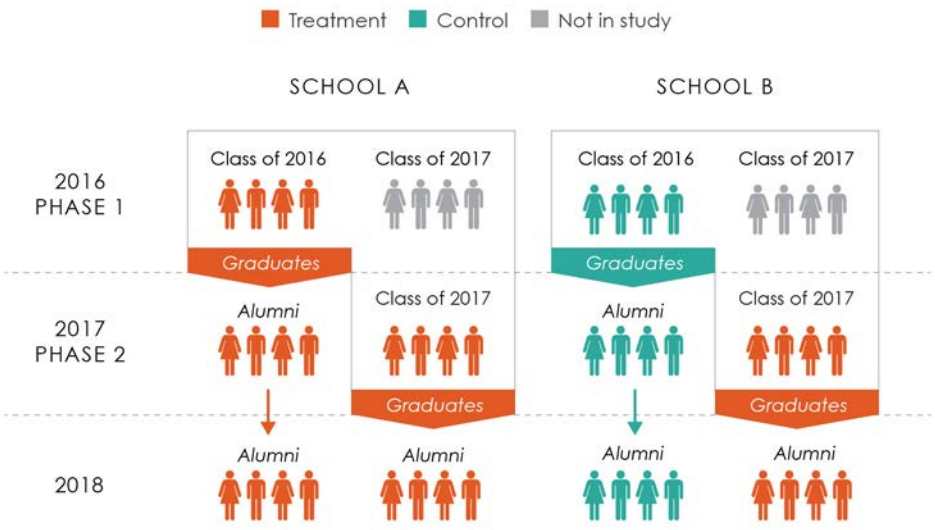

FIGURE 1.2: MOVING THROUGH THE UNIT OF RANDOMIZATION

Figure 1.2 illustrates an example where the unit of randomization is the school rather than the individual, the program is available for students in grade twelve, and phase one begins at the start of the 2016 school year while phase two is delayed until the subsequent school year. Assume that there are only two schools in the sample, School A and School B. Orange indicates that the given class received computers.

In phase one, all eligible students in grade twelve (i.e., the class of 2016) at School A receive a home computer. In phase two, all eligible students in grade twelve (i.e., the class of 2017) at both schools receive a home computer. Since phase two takes place one school year after phase one, researchers can determine the long-term impact of the program by following up with recent graduates (i.e., the class of 2016) from School B, assuming the program has non-academic effects that persist after graduation. These students are in grade twelve at the beginning of the program, but by the time the program reaches their school in phase two, they are graduates and do not receive the treatment. Therefore, they can serve as a pure control group. Note that this example is simplified; in reality, researchers would need to randomize more than two schools to have enough statistical power to detect the impact of the home computers.

- *If the program is phased in too quickly, researchers may not be able to detect program impacts.* In this case, suppose the effect of home computers takes two years to materialize. This lag may be due to an adjustment period, the time it takes for students to learn how to use the computer and incorporate it into their routine. If the computers are only expected to change academic outcomes after two years, researchers would not be able to

detect the program's impact by comparing the treatment and control groups' outcomes at the end of one year. To avoid this pitfall, researchers and practitioners should ensure that the time for the program to affect outcomes is shorter than the time between the first and last phase-in (GT 2013, 130). While this can be a challenge in itself, it might be possible to estimate the time it may take for the impacts to materialize by consulting relevant literature on similar interventions that target similar populations.

Despite these limitations, many programs are designed to roll out gradually due to initial capacity constraints. If the rollout can be randomized, a phase-in design is a good way to measure short-term, and in some cases, long-term impacts of a program.

Challenge #2: Program has strict eligibility criteria

Overview

Many public and private programs have strict eligibility criteria (e.g., an income threshold or a categorical requirement) that separate the eligible from the ineligible. In such cases, randomization among eligible applicants may not be the most appropriate research design because they already have full access to the program.

CHALLENGE
When evaluating a program with strict eligibility criteria, it may not be appropriate to randomly assign eligible individuals to the treatment or control group.

IMPLICATION
A control group cannot be constructed.

SOLUTION
Relax the eligibility criteria and randomize among the marginally ineligible.

A Real-World Example

Medicaid is an example of a program that has strict eligibility criteria. Nevertheless, J-PAL affiliates Katherine Baicker (Harvard) and Amy Finkelstein (MIT), in collaboration with a number of researchers, were able to conduct a randomized evaluation to determine the impact of Medicaid on a wide range of outcomes. Medicaid is a government-run program designed to provide health insurance to low-income and disabled Americans. Eligibility for the program is determined by several factors including a cutoff based on the value of an applicant's income and assets (Taubman et al. 2014, 263). Figure 2.1 illustrates the categorically eligible groups: children and pregnant women, the disabled, and families enrolled in Temporary Assistance to Needy Families (TANF). To construct a control group, it was neither legally possible nor ethically appropriate to randomly prevent some eligible individuals from enrolling in Medicaid. By law, if individuals submit an application and are eligible, they must be granted access to Medicaid.

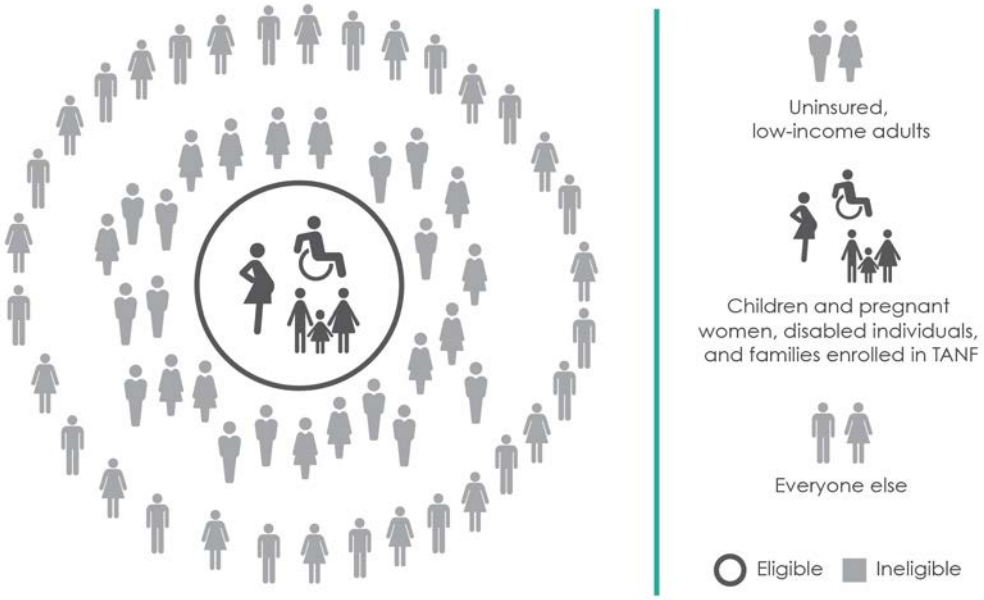

FIGURE 2.1: STRICT ELIGIBILITY CRITERIA

Solution: ***Randomize among the marginally ineligible***

To overcome this challenge, researchers took advantage of a Medicaid program expansion that was already scheduled to occur in 2008. The state of Oregon designed this expansion to provide 10,000 new spots to uninsured adults. The new spots were allotted by lottery to individuals who, before the program expansion in 2008, were not eligible for Medicaid benefits.

Expanding the eligibility criteria and then conducting a lottery among the newly eligible individuals allows researchers to conduct a randomized evaluation among this group, while maintaining access to the program for individuals who were already eligible prior to the study.

There are many ways to expand eligibility criteria. For example, the minimum income threshold could be relaxed or access to the program could be granted to an additional category of individuals. In the case of the Oregon Health Insurance Experiment, researchers took advantage of a categorical expansion. Medicaid in the state of Oregon continued to be available for children and pregnant women, the disabled, and families enrolled in TANF; however, in 2008, eligibility expanded to a new category of people: uninsured, low-income adults who were otherwise ineligible for Medicaid benefits (Finkelstein et al. 2012, 1062-1064).

To randomize among the marginally ineligible, as illustrated in Figure 2.2, first, expand the program's eligibility criteria.

FIGURE 2.2: EXPANSION OF ELIGIBILITY CRITERIA

This makes the program available to a group of individuals (e.g., uninsured, low-income adults in the case of the Oregon Health Insurance Experiment) who were ineligible prior to the expansion; these individuals are called the marginally ineligible.

Second, as illustrated in Figure 2.3, use a lottery to assign the marginally ineligible to the treatment and control groups.

FIGURE 2.3: LOTTERY AMONG THE MARGINALLY INELIGIBLE

When there are limited resources to serve all the marginally ineligible individuals, random assignment through a lottery can be used to fairly assign treatment. This approach is helpful because it does not require withholding the program from individuals who have been eligible in the past.

Limitations

- *Additional resources may be required for the implementer to accommodate more individuals.* In order to maintain the program for individuals who have always been eligible and expand the program to the marginally ineligible, the implementer will likely need access to more funding and personnel to run the program at a higher capacity (GT 2013, 124). Therefore, this approach is easiest to use when funding for a program expansion is already planned.
- *Randomizing among the marginally ineligible is an appropriate research design when the relevant policy question is: "Would this program be effective if it were extended to a new population?" but it may not be the best design if the question is: "Is this program effective for the currently eligible population?"* The marginally ineligible individuals may differ from the program's target population, in which case the program's impact on this population cannot be generalized to the currently eligible population. If the evaluation does not detect an impact on the marginally ineligible, it is still possible that the program has a positive effect on the currently eligible individuals—especially if individuals who are worse-off (and currently eligible) are more likely to benefit from the program. In the case of the Oregon Health Insurance Experiment, the state and researchers were particularly interested in evaluating whether Medicaid benefits were effective for the newly eligible population of uninsured, low-income adults. Therefore, the experimental design (i.e., randomization among the marginally ineligible) was an appropriate way to answer their policy-relevant question.

Challenge #3: Program is an entitlement

Overview

Many programs aimed at reducing poverty are entitlement programs (i.e., "government program[s] that [guarantee] certain benefits to a particular group or segment of the population" (Oxford 2017)). Oftentimes, these programs experience low take-up rates (i.e., despite open program slots, many eligible individuals are not enrolled). When evaluating a non-compulsory, entitlement program, researchers and practitioners cannot and should not force individuals to take up the program nor deny eligible individuals access to the program. Selecting a group of eligible individuals and randomizing access to the program would infringe on their right to receive the program.

CHALLENGE
When evaluating a non-compulsory, entitlement program, eligible individuals cannot and should not be forced to take up the program or denied access to the program.

IMPLICATION
Access to the program cannot be randomized to create a treatment group and a control group.

SOLUTION
Randomly distribute an encouragement to take up the program.

An Example

The Supplemental Nutrition Assistance Program (SNAP) provides nutrition assistance to eligible, low-income individuals and households in the United States. Randomly assigning access to SNAP benefits among eligible individuals to create a treatment and control group is not feasible because the program is an entitlement.

Solution: Encouragement design

Instead of randomly assigning access to SNAP benefits, researchers could use an encouragement design as illustrated in Figure 3.1.

An **encouragement design** is, "a research [design] in which both treatment and control groups have access to the program, but some individuals or groups are randomly assigned to receive encouragement to take up the program" (GT 2013, 445). Rather than randomly assigning access to the program, researchers randomly assign an encouragement. The encouragement can be a small incentive, letter, postcard, or phone call that reminds people of their eligibility and details steps to enroll in the program. Effective encouragement leads to higher take-up of the program in the treatment group than in the control group. It is important to note that it is the impact of receiving an encouragement to take up the program that is evaluated (and its indirect effect on program take-up), rather than the direct impact of the program itself.

FIGURE 3.1: ENCOURAGEMENT DESIGN

Since SNAP benefits cannot and should not be withheld from eligible individuals, researchers can use the encouragement design to create a control group in a different way. Both the treatment and control groups are comprised of potentially eligible individuals who currently are not enrolled in SNAP. Those in the treatment group receive specific encouragement to apply for SNAP benefits and those in the control group do not receive any specific encouragement.

The control group does not lose access to SNAP benefits as a result of the intervention. The intervention does not limit the usual encouragement and assistance that the individuals might encounter through other channels, and they can still apply for SNAP benefits at any time.

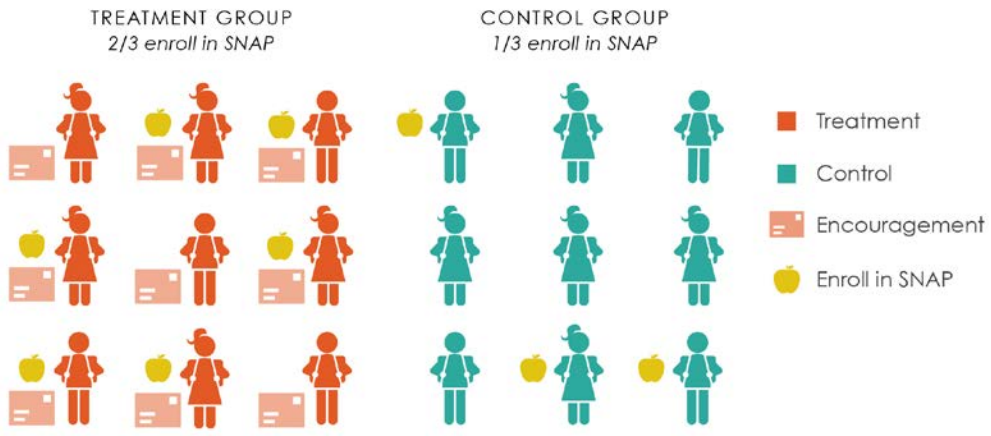

FIGURE 3.2: COMPARE THE ENTIRE TREATMENT GROUP TO THE ENTIRE CONTROL GROUP REGARDLESS OF ENROLLMENT STATUS

When studying the impact of SNAP benefits on key outcomes, such as spending, health care utilization, or nutrient intake, it is important to compare the *entire* treatment group to the *entire* control group. As illustrated in Figure 3.2, individuals in the treatment group who receive the encouragement but do not apply for SNAP benefits must still be considered a part of the treatment group when analyzing the results. Similarly, individuals in the control group who decide to apply for SNAP benefits without any special encouragement must remain in the control group for analysis.

Limitations

- *For an encouragement design to work, the program to be evaluated must be undersubscribed* (GT 2013, 135). If everyone who is eligible is already enrolled in SNAP, then there would not be enough unenrolled, eligible individuals for the study.
- Careful consideration of the design of the encouragement is important.
    - *To generate impact estimates, the encouragement must induce significantly higher take-up rates in the treatment group compared to the control group.* If 30 percent of treatment individuals apply for and are ultimately enrolled in SNAP and 28 percent of control individuals decide to apply and are ultimately enrolled in SNAP, this difference in take-up rates will likely not be sufficient to evaluate SNAP's impact. This is because in the analysis phase, receipt of encouragement is used as a proxy for program enrollment

(GT 2013, 135). When we compare an outcome, such as the number of diet-related hospitalizations,[1] we must compare the entire treatment group to the entire control group, regardless of which individuals in either group ultimately applied for or enrolled in SNAP. As such, we estimate the impact of the encouragement to apply for SNAP benefits. This comparison is an appropriate way to estimate the actual impact of SNAP benefits on diet-related hospitalizations only if a substantially larger portion of the treatment group applied for and enrolled in SNAP compared to the control group. Note the relationship between sample size and the difference in take-up rates: with a smaller sample size, the difference in take-up rates between the treatment group and control group must be large in order to detect the impact of a program. Conversely, with a larger sample size, the difference in take-up rates does not have to be quite as large to generate a precise impact estimate.

- *The encouragement should not have a direct effect on the outcome* (GT 2013, 135). For example, say researchers wanted to measure the impact of SNAP benefits on nutritional outcomes. If application assistance is provided in person, the encouragement may include a free round-trip subway pass to the application-assistance office. If the office is located next to a food pantry, treatment individuals may pick up food from the pantry after receiving application assistance. Impact estimates may indicate that treatment individuals are more nourished than control individuals. However, it would be impossible to distinguish the impact of the subway pass, which brought treatment individuals closer to a food pantry and thus more likely to pick up food on the way home, and the impact of SNAP benefits. Part of the encouragement, the subway pass, had a direct effect on the nutritional outcomes of treatment individuals.[2]
- *"Everyone must be affected by the encouragement incentive in the same direction"* (GT 2013, 137). For example, say the SNAP encouragement letter includes the following:

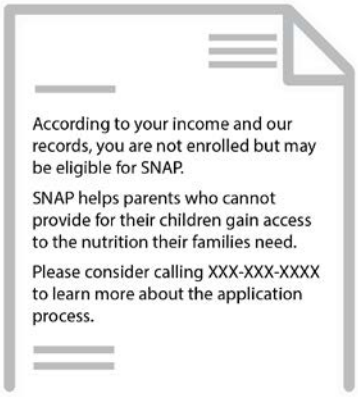

FIGURE 3.3: HYPOTHETICAL SNAP ENCOURAGEMENT LETTER

[1] For individuals with diabetes, food insecurity can lead to increased rates of hypoglycemia and subsequent hospitalizations. A hospitalization due to hypoglycemia is an example of a diet-related hospitalization. Evidence exists to suggest that SNAP reduces food insecurity, which may in turn reduce diet-related hospitalizations (U.S. Executive Office of the President 2015, 23-24).

[2] This example is adapted from the pepper farmers training program example originally found in GT 2013, 135-136.

> A mother may not respond well to this letter if she interprets the phrases "you are not enrolled" and "parents who cannot provide for their children" as negative or accusatory. Some parents may overlook these phrases, but others may be discouraged from applying for SNAP benefits due to the negative tone of the letter. "If the encouragement itself increases the take-up of some groups and reduces the take-up of others, [impact estimates will likely be biased]" (GT 2013, 137).

Setting the stage for an encouragement design

Researchers Amy Finkelstein (MIT) and Matthew Notowidigdo (Northwestern) are working with Benefits Data Trust (BDT), a not-for-profit organization based in Philadelphia, on a randomized evaluation to measure the effects of different outreach strategies on SNAP application and enrollment rates among potentially eligible individuals in Pennsylvania. BDT is administering the outreach, which includes a combination of letters, reminder postcards, and application assistance.

This evaluation is not considered an encouragement design because the sample size is only large enough to identify the impact of *outreach strategies* on the likelihood of applying for SNAP benefits (an immediate outcome) rather than the impact of *SNAP benefits* on individuals' health and wellbeing (subsequent outcomes). If this study were designed with a larger sample size, or if the outreach were to result in a substantially higher take-up rate in the treatment group, it could be used to measure the impact of SNAP benefits on subsequent outcomes such as individuals' health and well-being. Additionally, if the study were to have these elements (i.e., larger sample size and take-up rate), it could then be considered an encouragement design. In the meantime, findings from this evaluation could help identify an effective form of outreach that could be later used in a full-scale encouragement design to evaluate the impact of SNAP benefits on individuals.

Challenge #4: Sample size is small

Overview

Researchers and their implementing partners may not have enough resources to provide services to a large number of people, or there may not be enough eligible individuals to include in an evaluation. Conducting an evaluation with a small sample size may decrease the likelihood that the researchers will be able to detect an impact—even if the program is effective. More formally, a study is said to be "underpowered" when the sample size is too small to allow for detection of a reasonably sized impact, even if the program is effective. An underpowered randomized evaluation may consume substantial time and monetary resources and provide little useful information, or worse, tarnish the reputation of a (potentially effective) program. For more information, please refer to the J-PAL Research Resource: The Danger of Underpowered Evaluations.

CHALLENGE
The sample size is small.

IMPLICATION
With fewer randomized units, the study may not have enough statistical power to detect a program's impact—even if the program is effective.

SOLUTION
Randomize at a lower level and/or use stratified random assignment.

An example

Say researchers are designing an evaluation to determine the impact of an after-school tutoring program. The sample includes 400 students who are enrolled in twelve schools, and the unit of randomization is the school. As illustrated in Figure 4.1, students in six of the schools are randomly assigned to the tutoring program while the other six schools are assigned to the control group. Since the number of units randomly assigned to the treatment and control groups affects the power of the experiment, with six treatment units and six control units, the power of the experiment may not be sufficient to detect the program's impact.

FIGURE 4.1: SMALL SAMPLE SIZE

Solution 1: Randomize at a lower level

It may be possible to randomize at a lower level such as classrooms instead of entire schools. Suppose each school has two classrooms. As illustrated in Figure 4.2, instead of randomly assigning twelve schools to be in the treatment and control group, randomly assign the 24 classrooms into the treatment or control group. This increases the number of randomized units and increases the statistical power of the evaluation.

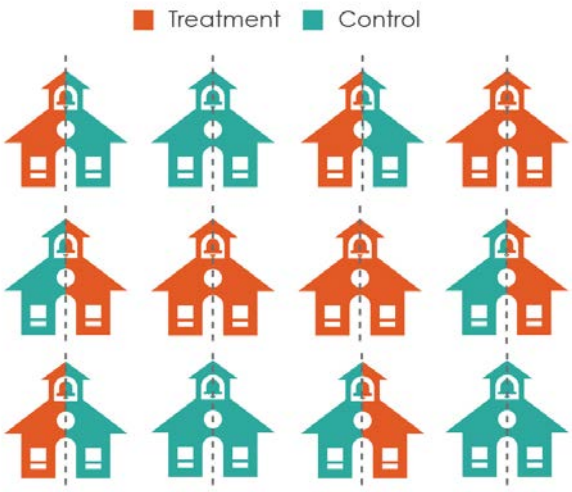

FIGURE 4.2: RANDOMIZATION AT THE CLASSROOM LEVEL

Limitation

- *Decreasing the level of randomization may increase the possibility of* ***spillover effects***. Illustrated in Figure 4.2, when the unit of randomization is the classroom, by random assignment some schools contain a mix of treatment and control classrooms, while other schools contain 100 percent treatment classrooms or 100 percent control classrooms. Spillovers are likely to occur in the schools with both a treatment and control classroom. For example, if students from treatment classrooms study with students from control classrooms, they may share lessons learned from the tutoring program with their control group peers. Thus, the control students may share the benefits of the program, and impact estimates might be underestimated. While spillover effects reduce the statistical power of an experiment, it may still be worthwhile to randomize at a lower level. In some cases, the benefits of increased power from randomizing at a lower level outweigh the costs of spillover effects.
- *Randomizing at the classroom level may incentivize teachers to shift perceived high-need students into treatment classrooms, which may compromise random assignment*. If class rosters are flexible, following random assignment, teachers may decide to transfer students who are perceived to be high-need into the classrooms that are assigned to the treatment group. This would change the composition of the treatment and control groups and may lead to an imbalance in the average academic performance level between the two groups. There are other randomization strategies that ensure all high-need students receive treatment (e.g., see Challenge #2 for **randomizing among the marginally ineligible**), but in this case, to avoid students shifting classrooms, random assignment could be conducted after class rosters have been established.

Solution 2: Stratified random assignment

Stratification is "an assignment method in which the sample is first divided into strata or groups based on observable characteristics, and then," within each group, individuals are randomly assigned to the treatment or control group (GT 2013, 451). This method ensures that there are balanced proportions of treatment and control students within each subgroup; while randomization achieves this balance *in expectation*, stratified random assignment *ensures* that treatment and control groups are equivalent on key observable variables. By doing so, the effect of the intervention is isolated, and one can more confidently attribute any differences in outcomes to the presence of the intervention. Additionally, "stratifying on variables that are strong predictors of the outcome can increase [the] statistical power [of the experiment]" and allows researchers to identify the program's impact on specific subgroups (GT 2013, 154).

Returning to the tutoring program, say the unit of randomization is the classroom. In the absence of stratification, the ratio of treatment and control classrooms may vary significantly from school to school. As illustrated in Figure 4.2, some schools have no treatment classrooms, others have one, and others have two treatment classrooms. To avoid this imbalance, researchers can stratify by school.

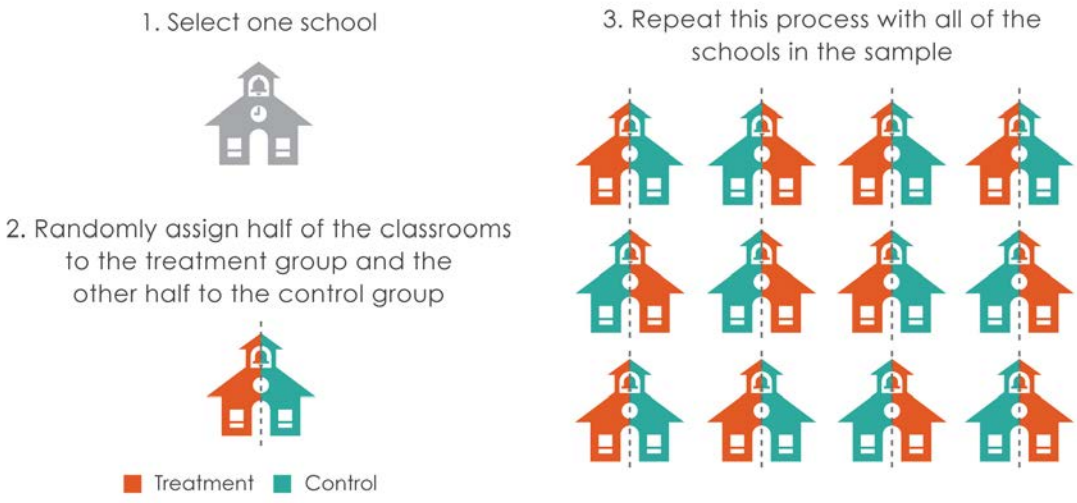

FIGURE 4.3: STRATIFICATION BY SCHOOL

With stratification, researchers can ensure that within each school, there is an equal ratio of treatment classrooms to control classrooms. This balance also helps researchers identify whether the after-school tutoring program has a different impact on students based on the school itself.

Additionally, from an operational standpoint, an equal proportion of treatment and control classrooms in each school may make it easier for service providers to implement the after-school tutoring program. Prior to stratification, service providers responsible for 100 percent treatment schools would need to serve significantly more classrooms compared to service providers responsible for 50 percent treatment schools. By standardizing the proportion of treated classrooms within each school, stratification distributes the implementing load more equally across service providers and reduces strain on particular service providers.

From a political standpoint, stratified random assignment is a fair way to allocate funding across schools. Prior to stratification, schools where 100 percent of classrooms were assigned treatment would receive significantly more funding compared to schools in which 50 percent or none of the classrooms were assigned treatment. The after-school tutoring program will likely be perceived as more fair if each school in the study is able to enroll half of their classrooms in the program.

Limitations

- *"Stratifying on variables that are not highly correlated with the outcome can [reduce statistical power]"* (GT 2013, 161). In the tutoring program example, say researchers stratified on the teacher's gender in addition to stratifying by school. If gender is not correlated with academic outcomes, stratifying on this variable could unnecessarily reduce statistical power.
- *Stratifying on too many variables may create unbalanced subgroups* (GT 2013, 157). The number of variables on which you can stratify is limited by the number of people in your sample that will fall into the subgroups. The more variables you stratify on, the smaller each subgroup becomes.

IMPLEMENTATION CHALLENGES

Challenge #5: It is difficult for service providers to adhere to random assignment due to logistical or political reasons

Overview

For proper program implementation, service providers must give treatment to the individuals assigned to the treatment group and leave individuals in the control group untreated. In some cases, especially when treatment and control individuals are in close proximity, this is a difficult task.

CHALLENGE
It is difficult for service providers to adhere to random assignment due to logistical or political reasons.

IMPLICATION
Administering the incorrect treatment can lead to crossovers and contamination of the control group.

SOLUTION
Assign treatment and control groups to different service providers and/or increase the level of randomization.

An Example

Consider a hypothetical scenario in which a health care provider wants to evaluate a new approach to increase take-up of annual, preventative health care visits by scheduling parent and child visits for the same date and in the same location. At the end of every visit with the doctor, all parents are given an opportunity to schedule the next appointment for their child. Additionally, parents in the treatment group are given an opportunity to book an appointment with their own physician, scheduled on the same day and at the same location as the child's next appointment. This consolidated check-up program aims to improve parental health by making it more convenient to see a doctor.

It might be logistically or politically difficult for service providers to distinguish between treatment and control groups. In this case, the nurse who concludes patients' appointments needs to offer consolidated check-ups to some parents and not others. If a few families are waiting to schedule their child's next appointment, it would be challenging for the nurse to offer the service to one family and not mention the service to the next family in line. Having one nurse serve families from both the treatment and control groups in close proximity may lead to **crossovers**, **contamination**, and ethical challenges.

Illustrated in Figure 5.1, a **crossover** would occur if a nurse offers the consolidated check-up service to a family in the control group because they are mistaken for a treatment group family, or a nurse knowingly offers the service to a control family because the nurse is sympathetic to a family in need and feels ethically obliged to make them aware of any service that may help. In either of these situations, the validity of the control group is compromised because some control families receive treatment.

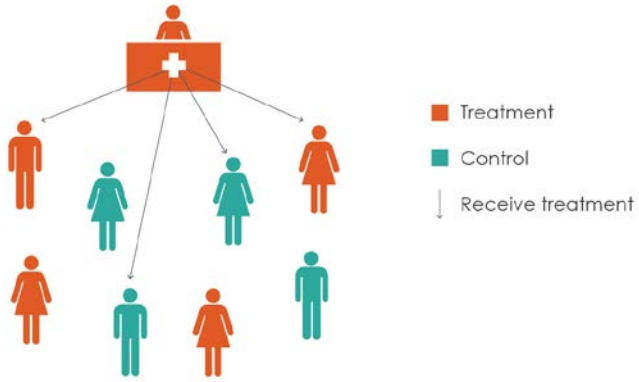

FIGURE 5.1: INCORRECT DISTRIBUTION OF TREATMENT

Solution 1: Assign treatment and control groups to different service providers

Illustrated in Figure 5.2, if service providers (i.e., nurses) within a clinic have trouble distinguishing between treatment and control groups and customizing the service for each group, in each clinic, train one group of nurses to give the treatment and train another group to administer the standard service for the control group. This way each nurse is only responsible for providing one type of service. At the end of a child's appointment, the family will be randomized into the treatment or control group. The appropriate nurse will meet with the family to offer to schedule the child's next appointment. If the family has been assigned to treatment, the treatment nurse will also offer to schedule the parent's appointment. Note that the unit of randomization is still the family rather than the nurse or the clinic.

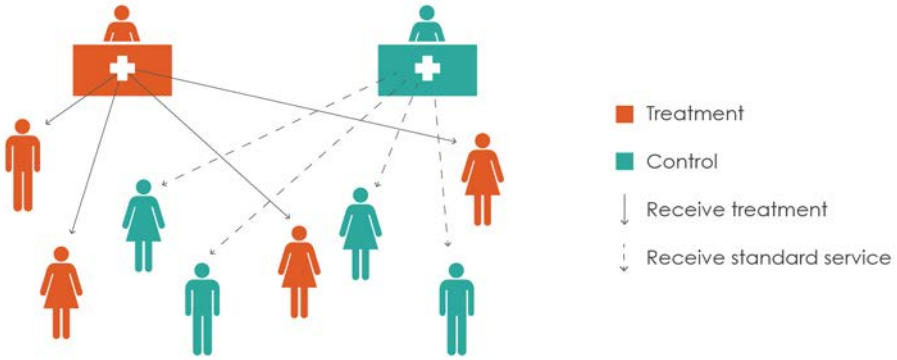

FIGURE 5.2: ASSIGN TREATMENT AND CONTROL GROUPS TO DIFFERENT SERVICE PROVIDERS

Limitations

- *If treatment and control individuals are in close proximity, this solution may be perceived as unfair.* By assigning treatment and control groups to different nurses, the nurses are more likely to administer the correct service. However, nurses may perceive this design as unfair because some families visiting the same clinic are not offered the same consolidated check-up benefit as the others.

- *To maintain statistical power, ensure that the treatment is implemented consistently across service providers.* Assuming that there is more than one nurse in each clinic that is trained to provide the treatment service (e.g., Nurse A and Nurse B), it is important that the method Nurse A uses to notify a treatment family of the consolidated check-up service is consistent with the method Nurse B uses. If delivery of the treatment is inconsistent across nurses, it will be difficult to isolate the impact of the program from the impact of the different service delivery styles of Nurse A and B.

Solution 2: Increase the level of randomization

It is also possible to increase the level of randomization. As illustrated in Figure 5.3, instead of randomizing at the family level, researchers could randomize at the clinic level. This way, every family that visits a given clinic will be offered the same services based on the clinic's treatment or control status. Instead of only training half of a clinic's nursing staff to provide the treatment, the treatment clinic's entire nursing staff will be trained to administer the treatment. This will reduce the ethical and logistical challenges of providing different services to different families in close proximity.

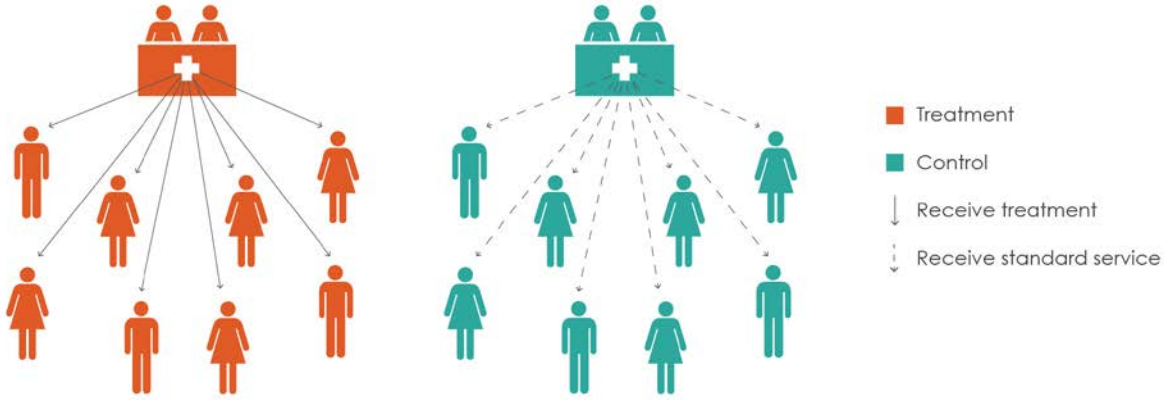

FIGURE 5.3: RANDOMIZATION AT THE CLINIC LEVEL INSTEAD OF THE FAMILY LEVEL

Limitation

- *If we increase the level of randomization, the number of randomized units decreases* (e.g., if we have 2 clinics with 100 families in each clinic, the number of randomized units falls from 200 to 2). If everything else in the experiment remains the same, randomizing at a higher level will decrease the experiment's statistical power. "This is because the outcomes of people in the same unit [i.e., the same clinic] are not fully independent of each other (GT 2013, 117)."
- *By randomizing at a higher level, individuals may self-select to receive services.* If researchers randomize at the clinic level and patients hear about the new program, they may choose to visit the clinic that offers the consolidated check-up program. In this case, the composition of the control and treatment groups would change, which may lead to an overestimation or an underestimation of the program's impact.

Challenge #6: Control group finds out about the treatment, benefits from the treatment, or is harmed by the treatment

Overview

Interaction between the control and treatment groups can lead to a number of challenges that threaten a randomized evaluation. The control group may find out about the treatment and react unfavorably, benefit from the treatment, or be harmed by the treatment. This can lead to spillovers, crossovers, and attrition, all of which can threaten an evaluation.

CHALLENGE
Control group finds out about the treatment and reacts unfavorably, benefits from the treatment, or is harmed by the treatment.

IMPLICATION
Spillovers, crossovers, and attrition can threaten the evaluation.

SOLUTION
Increase the unit of randomization and/or create a buffer.

A Real-World Example

An evaluation by J-PAL affiliates Bruno Crépon (*Centre de Recherche en Économie et Statistique*), Esther Duflo (MIT), Marc Gurgand (Paris School of Economics), Roland Rathelot (University of Warwick), and Philippe Zamora *(Centre de Recherche en Économie et Statistique)* studied the impact of career counseling on outcomes for young, college-educated job seekers in France. The evaluation covered 57,000 job seekers in 235 labor markets (Crépon et al. 2013). Researchers wanted to know: how do intensive job counseling services affect employment rates among those who receive counseling and in the overall job market? Researchers randomly assigned unemployed individuals to receive job counseling.

Spillovers, **crossovers**, and **attrition** may occur as a result of the following scenarios:

1. If individuals from the treatment and control group know each other, the control group may find out about the treatment and get upset because they are not receiving treatment. Service providers may lose the support of the community, and the control group might become unwilling to participate in the study, which could lead to **attrition**.
2. The control group may benefit from the treatment in a few ways that can lead to **spillovers** and **crossovers**.
    - *The treatment group may share its benefits with the control group.*

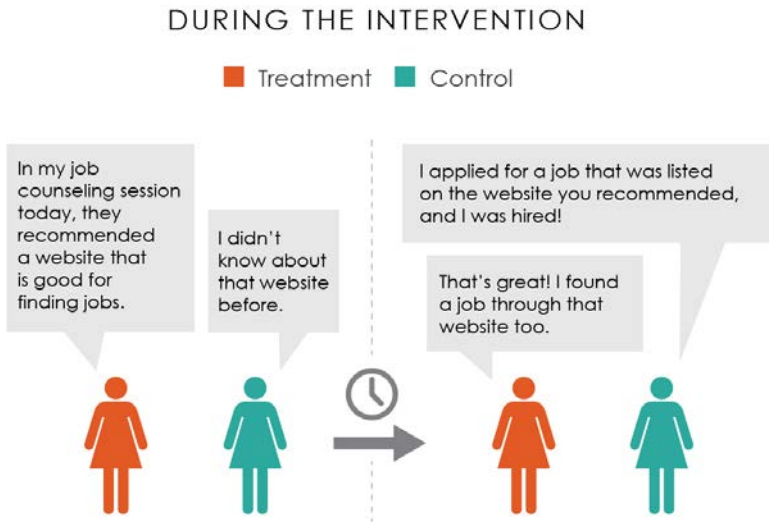

FIGURE 6.1: TREATMENT SHARING BENEFITS WITH THE CONTROL GROUP

As illustrated in Figure 6.1, if treatment individuals have close friends in the control group, they may share job-finding strategies from the job counseling sessions with their friends in the control group. If control individuals benefit from the information sharing and are as likely as treatment individuals to get jobs, researchers may conclude that the job counseling program had no effect. However, in reality impact estimates are likely underestimated due to the positive **spillover** effect.

- *The control group can benefit from treatment because they change their behavior*. Job seekers in the control group might notice that their peers are receiving job counseling and seek job counseling themselves from a different organization. As a result, the control group will not adequately represent what would have happened without job counseling: the impacts of the program may be underestimated due to this positive **spillover**. If a control group individual discovered a way to receive job counseling from the same service provider as their treatment group peers, this would be considered a **crossover**.

3. The control group may be harmed by the treatment if treatment and control individuals compete with each other. Illustrated in Figure 6.2, this study dealt with the possibility of a "displacement effect," which is when job opportunities are transferred from individuals who do not receive counseling to those who do. A displacement effect is a negative **spillover**. As a result of the job counseling services, individuals in the control group may be harmed because they now face increased competition in the labor market for a limited number of jobs. If the evaluation indicates that the employment rate in the treatment group is higher than the employment rate in the control group, it may be the case that the intervention simply shifted the set of individuals who received the limited available jobs. As depicted in Figure 6.2, the total employment rate in the given labor market is fifty percent whether or not the experiment occurs. However, with the experiment, firms only hire treatment individuals, whereas without the experiment, firms hire a mix of individuals.

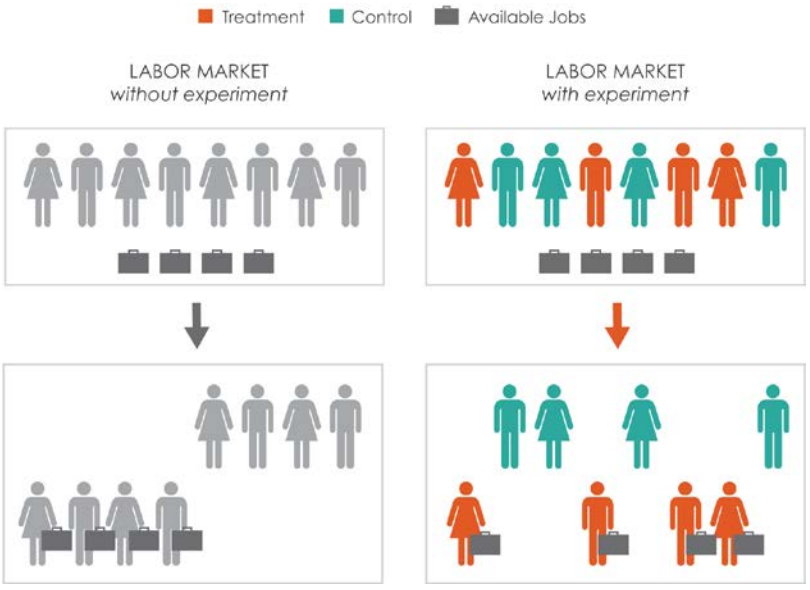

FIGURE 6.2: NEGATIVE SPILLOVER

In this case, job training helped job-seekers get the limited number of jobs, but it did so at the expense of individuals in the control group. This is an example of a displacement effect, which is a type of negative **spillover**. As a result of the displacement effect, the control group is harmed by the treatment, and the impact of the program may be overestimated.

Solution 1: Increase the level of randomization

To decrease the likelihood of interaction that leads to information sharing and behavioral changes between the treatment and control group, researchers can randomize at a higher level. Instead of the individual level, randomization can be implemented at the neighborhood or municipality level. "Choosing the level of randomization so that the most relevant interactions occur between people in the same group is the best way to limit spillovers" (GT 2013, 114).

Limitation

- *If we increase the level of randomization, the number of randomized units decreases* (e.g., If we have 4 labor markets, each with 100 unemployed individuals, the number of randomized units falls from 400 to 4). If everything else in the experiment remains the same, randomizing at a higher level will decrease the experiment's statistical power. "This is because the outcomes of people in the same unit [i.e., the same labor market] are not fully independent of each other" (GT 2013, 117).

Solution 2: Create a buffer to contain spillovers

When individuals or groups of individuals surrounding the treatment and control units are not included in the sample, they are considered buffers. Buffers decrease the likelihood of interaction between the treatment and control group. In the job counseling example, researchers could randomly assign treatment status to one labor market, leave the adjacent labor markets out of the study, and then randomly assign control status to a labor market a certain distance away from the treatment area to avoid spillovers. This is illustrated in Figure 6.3.

FIGURE 6.3: BUFFERS TO REDUCE SPILLOVERS AND CONTAMINATION

Limitation

- This solution may require partnering with a larger service provider because individuals in the study are spread over a larger area.

Instead of eliminating spillovers, design an evaluation to measure them

If the individuals in the buffer are included in the study sample, the researcher can capture interesting information about the spillover effects of the program. Rather than attempting to eliminate spillover effects, researchers conducting this study were particularly interested in designing a randomized evaluation that could measure the spillover effects of the job counseling services. To understand different spillover effects, such as displacement effects, they randomized the proportion of unemployed individuals in each treatment area that was offered counseling. The randomly assigned proportions of treated individuals are illustrated in Figure 6.4.

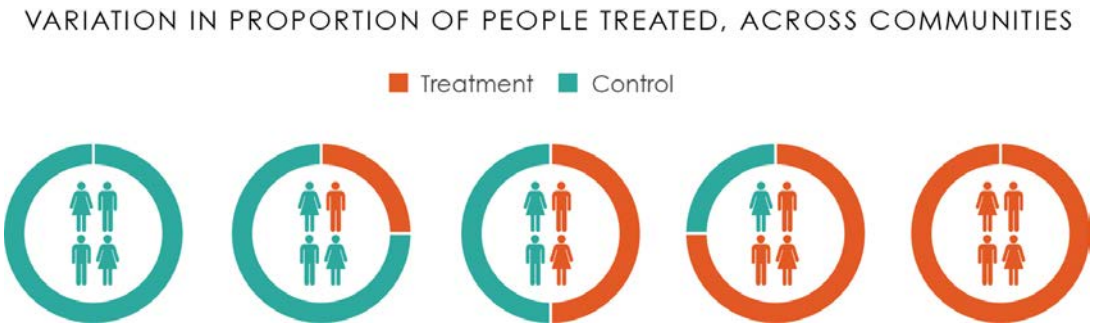

FIGURE 6.4: VARYING THE TREATMENT DENSITY

This is called varying the **treatment density**. As a result, researchers were able to detect whether the control group individuals in areas where they had to compete with a high proportion of counseled individuals (i.e., the 75 percent treated areas) were worse off than the control group individuals in areas with a low proportion of counseled individuals (i.e., the 25 percent treatment areas) (GT 2013, 179).

SUMMARY TABLE

CHALLENGE	IMPLICATION	SOLUTION	LIMITATIONS
Resources exist to extend the program to everyone in the study area	If everyone is treated at once, there is no control group	Use phase-in design	Anticipatory effects Difficult to evaluate long-term impacts If the program is phased in too quickly, researchers may not be able to detect program impacts
Program has strict eligibility criteria	May not be appropriate to randomly assign eligible individuals to the control group	Relax the eligibility criteria, and randomize among the marginally ineligible	Can only assess the impact on marginally ineligible individuals May need additional resources to accommodate more individuals
Program is an entitlement, so eligible individuals cannot be forced to take up the program or denied access to the program	Access to the program cannot be randomized to create a treatment group and a control group	Use encouragement design	Program must be undersubscribed The encouragement must increase take-up but be designed such that it influences nothing else
Sample size is small	Insufficient statistical power	Decrease the level of randomization Stratify on variables highly correlated with the outcomes	Decreasing the level of randomization may increase spillover effects Stratifying on too many variables may create unbalanced subgroups
It is difficult for service providers to adhere to random assignment due to logistical or political reasons	Crossovers Contamination	Assign treatment and control groups to different service providers Increase the level of randomization	May not be perceived as fair if different service providers are in close proximity Ensure that program is implemented consistently across service providers Increasing the level of randomization reduces statistical power
Control group finds out about the treatment and reacts unfavorably, benefits from the treatment, or is harmed by the treatment	Spillovers Crossovers Attrition	Increase the level of randomization Create a buffer	Increasing the level of randomization reduces statistical power May require partnering with a larger service provider

GLOSSARY

Anticipatory effect: "A change in behavior of a [control] group because they expect to receive access to the program later on (or in a rotation design where those in the current treatment group change their behavior because they know they are going to become the [control] group later)" (GT 2013, 443).

Attrition: When individuals drop out of the control or treatment group over the course of the evaluation.

Contamination: A result of crossovers. When a fraction of the control group has received treatment.

Counterfactual: What would have happened to the participants in a program had they not received the intervention. The counterfactual cannot be observed from the treatment group; it can only be inferred from the control group.

Crossover: When an individual in the control group strays from his or her initial assignment and receives the treatment.

Encouragement design: "A research [design] in which both treatment and control groups have access to the program, but some individuals or groups are randomly assigned to receive encouragement to take up the program" (GT 2013, 445).

Randomization among the marginally ineligible: A method that randomizes individuals that fall between an old threshold and a new threshold. All previously eligible individuals remain eligible for the program while a group of previously ineligible individuals becomes eligible for treatment.

Spillover: When a treatment affects those in the control group or individuals who are not in the study sample. Spillovers can take many forms and be positive or negative.

Statistical power: If a program has an impact, the likelihood that one's evaluation will detect this impact is given by the statistical power of the evaluation.

Stratification: An assignment method in which the sample is first divided into groups based on observable characteristics, and then, within each group, individuals are randomly assigned to the treatment or control group. For example, sample could be stratified based on gender, ethnicity, or age.

Treatment density: "The proportion of units (e.g., individuals, schools, or communities) within a geographic area that receive the treatment" (GT 2013, 178). Researchers can vary the treatment density to capture spillover effects.

REFERENCES

Glennerster, Rachel, and Kudzai Takavarasha. 2013. *Running Randomized Evaluations: A Practical Guide*. Princeton: Princeton University Press.

Crépon, Bruno, Esther Duflo, Marc Gurgand, Roland Rathelot, and Philippe Zamora. 2013. "Do Labor Market Policies Have Displacement Effects? Evidence from a Clustered Randomized Experiment." *The Quarterly Journal of Economics* 128(2): 531-80. doi: 10.1093/qje/qjt001.

Fairlie, Robert, and Jonathan Robinson. 2013. "Experimental Evidence on the Effects of Home Computers on Academic Achievement among Schoolchildren." *American Economic Journal: Applied Economics* 5(3): 211-40. doi: 10.1257/app.5.3.211.

Finkelstein, Amy, 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." *The Quarterly Journal of Economics* 127(3): 1057-1106. doi:10.1093/qje/qjs020.

Linden, Leigh L., Carla Herrera, and Jean Baldwin Grossman. 2013. "Achieving Academic Success After School: A Randomized Evaluation of the Higher Achievement Program." Working Paper. University Texas at Austin, July. http://www.leighlinden.com/Higher_Achievement.pdf.

Oxford Dictionary (US). 2017. "Entitlement Program." Accessed February 22. https://en.oxforddictionaries.com/definition/us/entitlement_program.

Taubman, Sarah L., Heidi L. Allen, Bill J. Wright, Katherine Baicker, and Amy N. Finkelstein. 2014. "Medicaid Increases Emergency-Department Use: Evidence from Oregon's Health Insurance Experiment." *Science* 343(6168): 263-68. doi: 10.1126/science.1246183.

U.S. Executive Office of the President. Council of Economic Advisers. 2015. "Long-Term Benefits of the Supplemental Nutrition Assistance Program." Accessed January 2017.