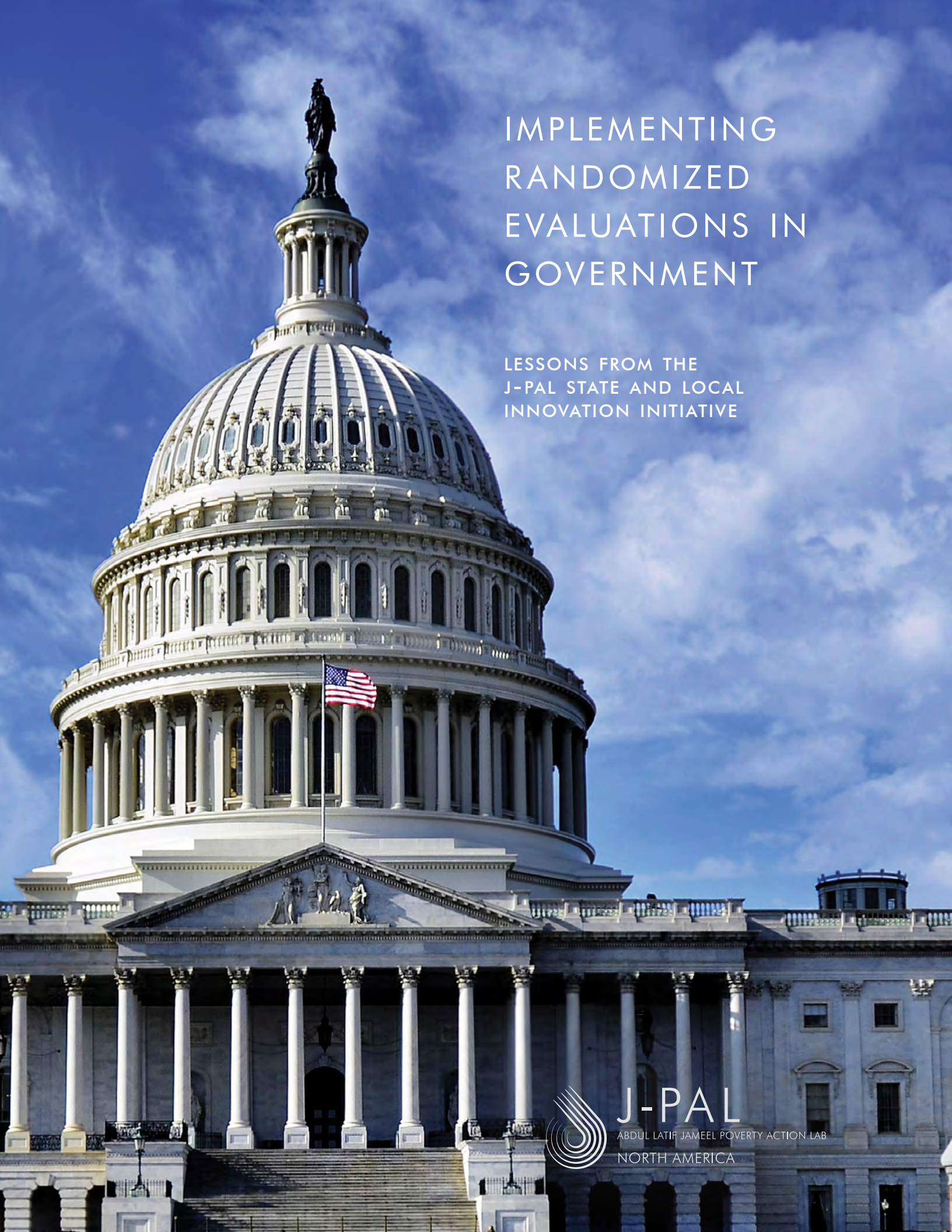


IMPLEMENTING RANDOMIZED EVALUATIONS IN GOVERNMENT

LESSONS FROM THE J-PAL STATE AND LOCAL INNOVATION INITIATIVE

J-PAL
ABDUL LATIF JAMEEL POVERTY ACTION LAB
NORTH AMERICA

Implementing Randomized Evaluations In Government:
Lessons from the J-PAL State and Local Innovation Initiative

Written by: Julia Chabrier, Todd Hall, Ben Struhl

J-PAL North America | September 2017 | povertyactionlab.org/na

COVER PHOTO: SHUTTERSTOCK.COM

SUMMARY

The J-PAL State and Local Innovation Initiative supports US state and local governments in using randomized evaluations to generate new and widely applicable lessons about the effectiveness of their programs and policies. Drawing upon the experience of the state and local governments selected to participate in the initiative to date, this guide provides practical guidance on how to identify good opportunities for randomized evaluations, how randomized evaluations can be feasibly embedded into the implementation of a program or policy, and how to overcome some of the common challenges in designing and carrying out randomized evaluations.

Please send comments, questions, or feedback to stateandlocal@povertyactionlab.org.

ACKNOWLEDGEMENTS

Many people provided guidance and advice in creating this document, for which we are grateful. First and foremost, we owe a special thanks to the state and local governments that have participated in this initiative for leading the projects discussed in this guide and for providing valuable feedback. We thank Mary Ann Bates, Jon Guryan and Melissa Kearney for providing valuable feedback and oversight for both the guide and for the initiative overall. Hannah Myers copyedited this document, and Amanda Kohn designed and formatted the final product. Claire Mancuso assisted in production of the guide as well. This work was made possible by support from the Alfred P. Sloan Foundation and the Laura and John Arnold Foundation. Any errors are our own.

ABOUT J-PAL NORTH AMERICA

J-PAL North America is a regional office of the Abdul Latif Jameel Poverty Action Lab (J-PAL), a global network of researchers who use randomized evaluations to answer critical policy questions in the fight against poverty. J-PAL's mission is to reduce poverty by ensuring that policy is informed by scientific evidence.

Founded at the Massachusetts Institute of Technology (MIT) in 2013, J-PAL North America leverages scholarship from more than 145 affiliated professors from over 40 universities, and a full-time staff of nearly 30 researchers, policy experts, and administrative professionals, to generate and disseminate rigorous evidence about the effectiveness of various anti-poverty programs and policies.

To address the complex causes and consequences of poverty, J-PAL North America's work spans a range of sectors including health care, housing, criminal justice, education, and labor markets.

ABOUT THE J-PAL STATE AND LOCAL INNOVATION INITIATIVE

The J-PAL State and Local Innovation Initiative supports US state and local governments in using randomized evaluations to measure the effects of programs and policies serving poor and vulnerable populations. The work of the initiative is aimed at enabling leaders within and beyond government to draw on evidence to support programs that work.

Through the J-PAL State and Local Innovation Initiative, J-PAL North America works to:

- Equip state and local governments with the tools to generate and use rigorous evidence;
- Share this evidence with other jurisdictions that may be facing similar challenges; and
- Document and disseminate best practices for feasibly implementing randomized evaluations at the state and local level.

The leaders selected to participate in this initiative work together to serve as models for others across the United States, demonstrating how state and local governments can create and use rigorous evidence to address challenging social problems.

INTRODUCTION

Since launching the J-PAL State and Local Innovation Initiative in 2015, J-PAL North America has received more than 50 letters of interest from state and local governments across the country. This initiative formed the basis for many conversations, conferences, training courses, and opportunities for mutual learning with these government partners. We have launched in-depth partnerships with eight state and local governments to develop randomized evaluations designed to inform their priority policy questions.

With this guide, we aim to share what we have learned from our partnerships with the governments that have participated in the initiative to date, so that other governments that are interested in pursuing randomized evaluations can learn from their experience. This guide also draws upon the experience of J-PAL's staff, who have worked with many different government agencies, nonprofits, and other partners, as well as the more than 800 ongoing and completed randomized evaluations conducted by J-PAL's affiliated researchers worldwide.

We provide practical guidance on how to identify good opportunities for randomized evaluations, how to embed randomized evaluations into program or policy implementation, and how to overcome some of the common challenges in designing and carrying out randomized evaluations. We also include links to resources and toolkits with more information. While some of the concepts in this guide are specific to randomized evaluations, many are applicable to other methods of impact evaluation as well.

The guide is organized into six sections:

- Why we launched the J-PAL State and Local Innovation Initiative
- What is a randomized evaluation and why randomize?
- Laying the groundwork for a research project
- Identifying opportunities for randomized evaluations
- Implementing an evaluation
- Making future evaluations easier

TABLE OF CONTENTS

Introduction iii
Why we launched the J-PAL State and Local Innovation Initiative 1
Lessons from the first year 2
What is a randomized evaluation and why randomize? 3
Common concerns about randomized evaluations 4
Beyond randomized evaluations 4
Laying the groundwork for a research project 5
Leveraging outside opportunities 5
Commitments from the government 5
Working with academic researchers 6
What does a successful research partnership look like? 7
Identifying opportunities for randomized evaluation 8
Defining a research question 8
Different ways to randomize 9
When does a randomized evaluation make sense? 9
Case Study: Philadelphia 11
How large a sample size would an evaluation need? 13
When not to do a randomized evaluation 14
Case Study: Pennsylvania 15
Defining and measuring outcomes 17
Implementing an evaluation 19
Ethics and Institutional Review Boards 19
Working with service providers and other stakeholders 19
The importance of piloting 20
Case Study: Rochester 21
Managing expectations around communication 23
Case Study: Puerto Rico 24
Conclusion: Making future evaluations easier 26
Strengthening administrative data systems 26
Changing the way research is framed 27
Building momentum 27
Case Study: South Carolina 28
Appendix 30
Bibliography 31

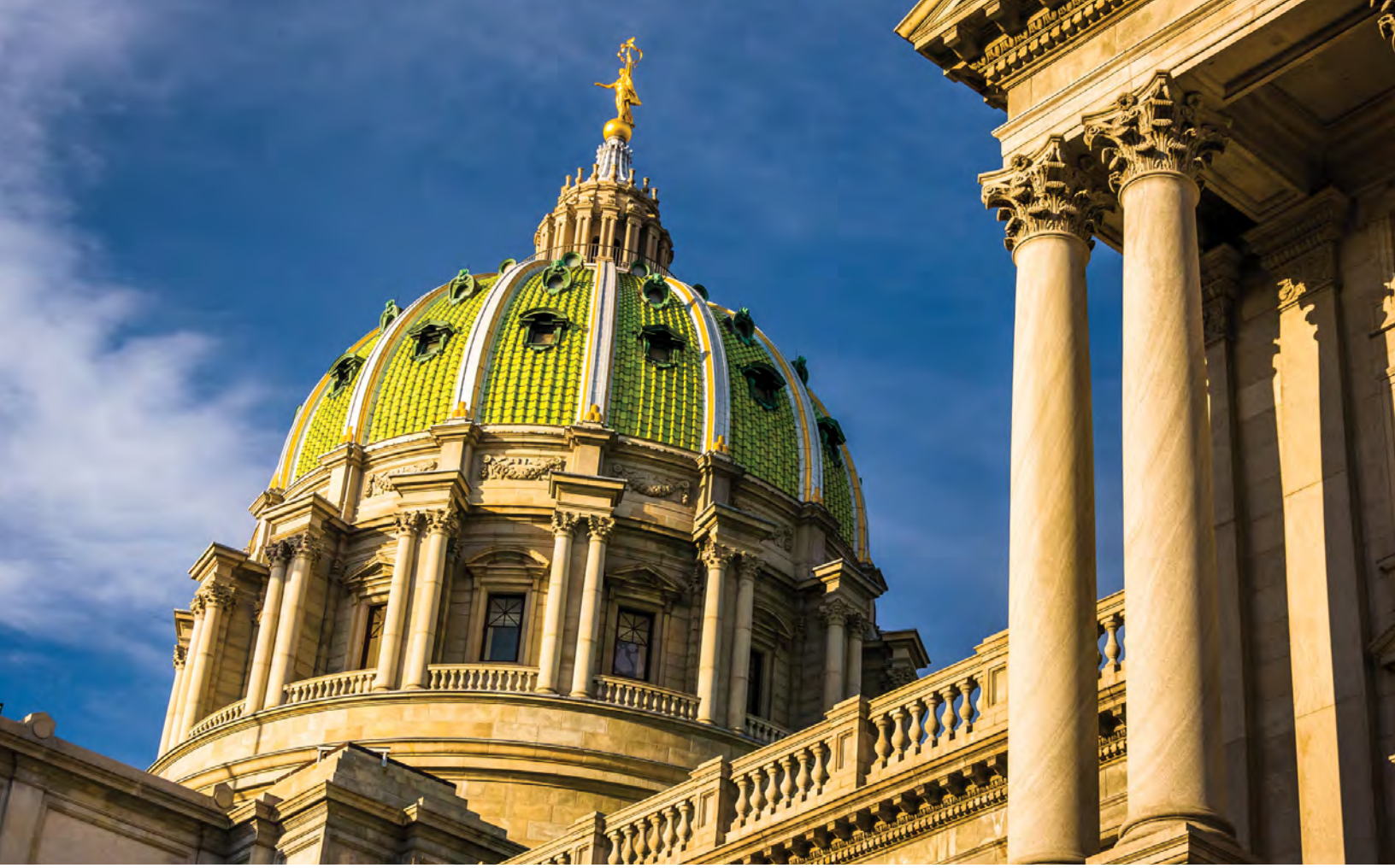


PHOTO: SHUTTERSTOCK.COM

WHY WE LAUNCHED THE J-PAL STATE AND LOCAL INNOVATION INITIATIVE

State and local governments across the United States are developing innovative solutions to address complex policy challenges, almost always with limited resources. Too often, they must make policy decisions without the benefit of rigorous evidence about what has been tried and proven elsewhere, or the opportunity to learn which of their own policies and programs are effective.

Randomized evaluations (also known as randomized controlled trials or RCTs) can be a powerful tool for generating rigorous evidence about the effectiveness of policies and programs. However, to date relatively few state and local governments have launched randomized evaluations.

There are a number of potential barriers to greater adoption of randomized evaluations by state and local governments. We suspect that many state and local policymakers are unsure how to identify opportunities to build randomized evaluations into their policies and programs. Policymakers often may not have connections with trusted and experienced researchers who can design a high-quality randomized evaluation. Another possible obstacle is a lack of early funding to launch "demonstration" evaluations.

We created the J-PAL State and Local Innovation Initiative to help jurisdictions address these barriers. The initiative features a two-phase competition for state and local governments interested in using randomized evaluations to inform their decision-making. In Phase I, selected state and local governments receive flexible pilot funding, connections with experienced researchers from J-PAL's network, and ongoing technical support from J-PAL North America staff to help them develop randomized evaluations. In Phase II, state and local governments can apply in

partnership with a researcher from J-PAL's network for additional funding to implement the evaluation. The goal of the State and Local Innovation Initiative is to generate evidence that state and local governments can use to improve their programs and policies and ultimately the lives of the people they serve. Randomized evaluations can have an impact far beyond the original context in which they were conducted when policymakers use rigorous evidence to improve or scale up programs with demonstrated success.

LESSONS FROM THE FIRST YEAR

There is untapped demand among state and local governments for high-quality evidence. Since launching the State and Local Innovation Initiative in 2015, J-PAL North America has received more than 50 letters of interest from city, county, and state governments across the United States. These letters of interest have proposed randomized evaluations to inform a wide range of policy issues, including crime and violence prevention, education, employment, health care, and homelessness. In some cases, our initial engagement with a government partner has revealed a wide appetite for evidence-based policymaking, sparking multiple efforts to generate or apply evidence across different government agencies.

State and local governments can be excellent partners for policy-relevant evaluations. Given their deep knowledge of the local context, state and local policymakers are well-positioned to identify which policies and programs would benefit most from rigorous evaluation, what ethical and feasibility considerations are important for a given study, and how to build support from key stakeholders. With financial and technical support from J-PAL North America and in partnership with researchers from J-PAL's network, the five governments selected in the first round of the State and Local Innovation Initiative have to date successfully launched three randomized evaluations, with a fourth in development. These randomized evaluations can serve as models for other governments, demonstrating that randomized evaluations may be possible in their context as well.

Governments across the country face many similar challenges and can share knowledge about what works to address them. Through the State and Local Innovation Initiative, governments have come forward with proposals to address critical challenges confronting state and local jurisdictions across the United States. For example, in the first round of the competition, multiple governments applied to develop evaluations of programs related to opioid and other substance use disorders. Several of these governments then participated in a conference hosted by J-PAL North America to brainstorm ways to test approaches to combat the opioid epidemic with other policymakers, researchers from J-PAL's network, and medical experts. Our work with state and local governments on this issue also informed a policy brief on strategies to combat the opioid epidemic, which we created at the request of the White House Office of National Drug Control Policy.

In the most recent round of the competition, preventing homelessness featured as a top policy concern among governments. We chose to partner with three governments to design evaluations of their innovative homelessness prevention programs, and plan to work with these governments and their research partners to share knowledge across sites. Ultimately, we aim to share what these governments learn about which approaches are most effective with the broader community of policymakers and researchers working to address this issue.

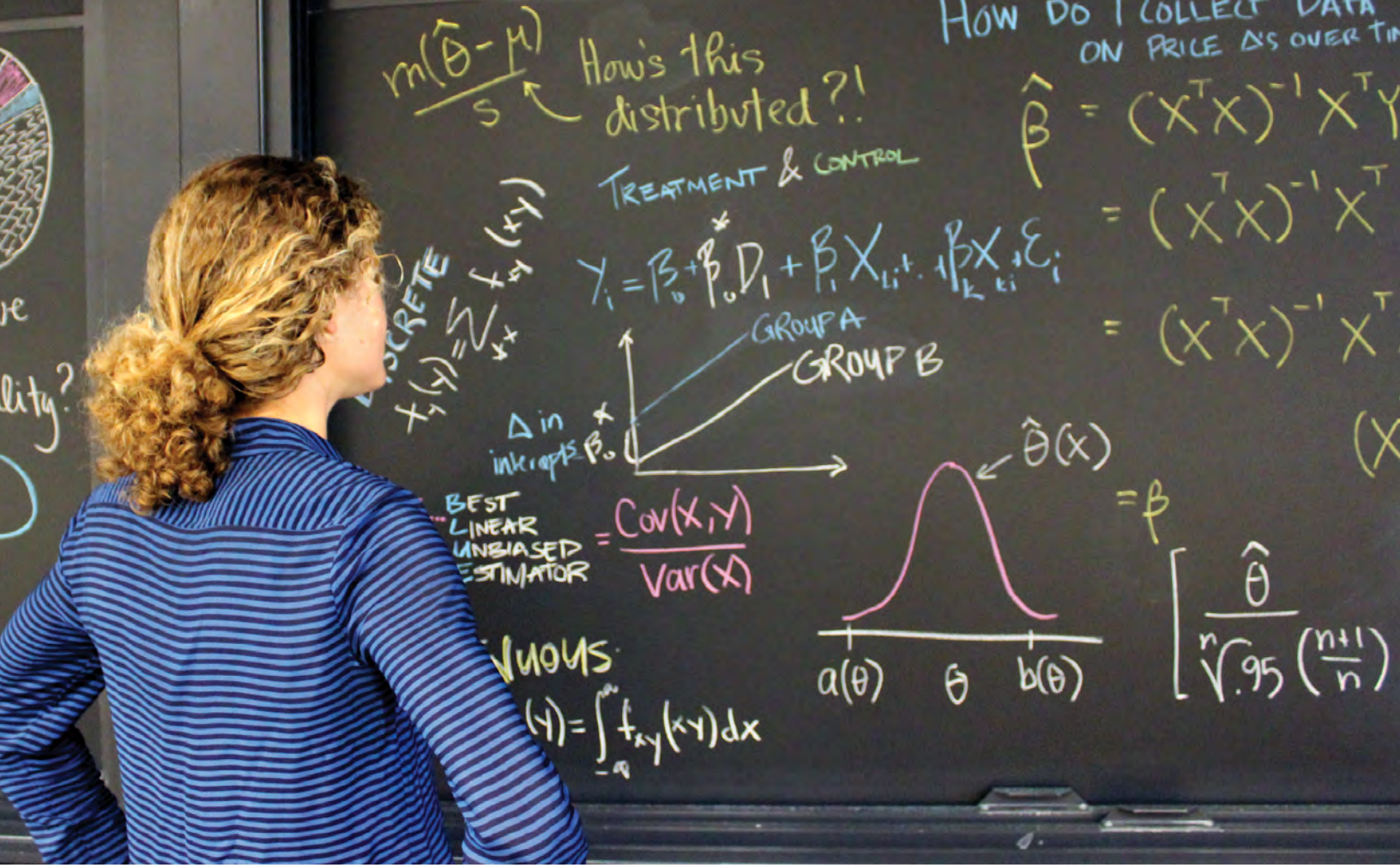


PHOTO: FRANCINE LOZA | J-PAL

WHAT IS A RANDOMIZED EVALUATION AND WHY RANDOMIZE?

A randomized evaluation is a form of impact evaluation. Like other forms of impact evaluation, the purpose of a randomized evaluation is to evaluate the impact of a program by comparing what happened to program participants to *what would have happened* to these same participants without the program.

Measuring outcomes for program participants is often straightforward. However, it is impossible to determine directly what would have happened to these same participants had they not taken part in the program. Instead, impact evaluations use different methods to identify a "comparison group" of non-participants who closely resemble the participants. We can estimate the impact of a program by comparing the comparison group's outcomes to those of the program participants. For example, the comparison group could include participants themselves before the program, or people who were eligible for the program but did not take part. The more closely the comparison group mirrors the participants, the more confident we can be that any differences in outcomes between the comparison group and the participants are due to the program.

In the most simple form of a randomized evaluation, people are randomly assigned either to a treatment group, which is eligible to participate in the program, or to a control group, which is not eligible to participate. Random assignment ensures that, with a large enough number of people, the two groups will be similar on average before the start of the program across both measurable characteristics (such as age, race, or household income) and immeasurable characteristics (such as personal motivation or social support). After the program is delivered, we then measure the outcomes of individuals in the two groups. Because

the two groups were similar on average before the start of the program, we can be especially confident that any later differences in outcomes are a result of the program.

Because randomized evaluations rely on fewer assumptions than other forms of impact evaluation, their results are less vulnerable to methodological debates, easier to explain, and are often more convincing to program funders or other audiences.

Read more: Why Randomize? A one page primer on randomized evaluations | bit.ly/2vlsXds

COMMON CONCERNS ABOUT RANDOMIZED EVALUATIONS

Is it ethical to assign people to a control group and deny them access to a program?

If we have rigorous evidence that a program is effective and enough resources to serve everyone who is eligible, then it would not be ethical to deny some people access to the program in order to conduct a randomized evaluation. However, in many cases, we do not know yet whether a program is effective. And, unfortunately, it is often the case that there are many more people who could benefit from a program than there are resources to serve. In that circumstance, a randomized evaluation may change how people enroll in a program, but not reduce the number of people that the program serves.

Aren't randomized evaluations very expensive?

Randomized evaluations are not inherently more expensive than other types of evaluations. The cost of an evaluation often depends on whether the evaluation is using original data (such as data collected through surveys) or administrative data, which are information collected, used, and stored primarily for purposes other than research. Evaluations that draw on existing administrative data generally are much lower cost than evaluations that generate new data by conducting surveys.

Is it possible to conduct a randomized evaluation without waiting years for the results?

The length of time required to measure the impact of a program is largely dependent on the outcomes that one is interested in measuring, rather than the evaluation method. For example, an evaluation designed to measure the impact of an early childhood education program on high school graduation rates would necessarily take longer to yield results than an evaluation designed to measure the impact of the same program on third-grade reading scores.

Read more: Common questions and concerns about randomization | bit.ly/2im23Rm

BEYOND RANDOMIZED EVALUATIONS

As mentioned previously, randomized evaluations are just one of many impact evaluation tools. They are not always the best tool in a given situation. While J-PAL focuses on randomized evaluations, many of the topics discussed in this guide are relevant to other research methods as well. Which evaluation method is best for the reader of this guide will depend heavily on the context in which he or she is working. Collaborating with a research partner can often be useful to help determine which approach will work best for a specific program and context. We aim to be transparent with our partners and recommend against doing a randomized evaluation in cases where another evaluation method would be a better fit.

Read more: Impact evaluation methods: what are they and what assumptions must hold for each to be valid? | bit.ly/2rZ2O2p

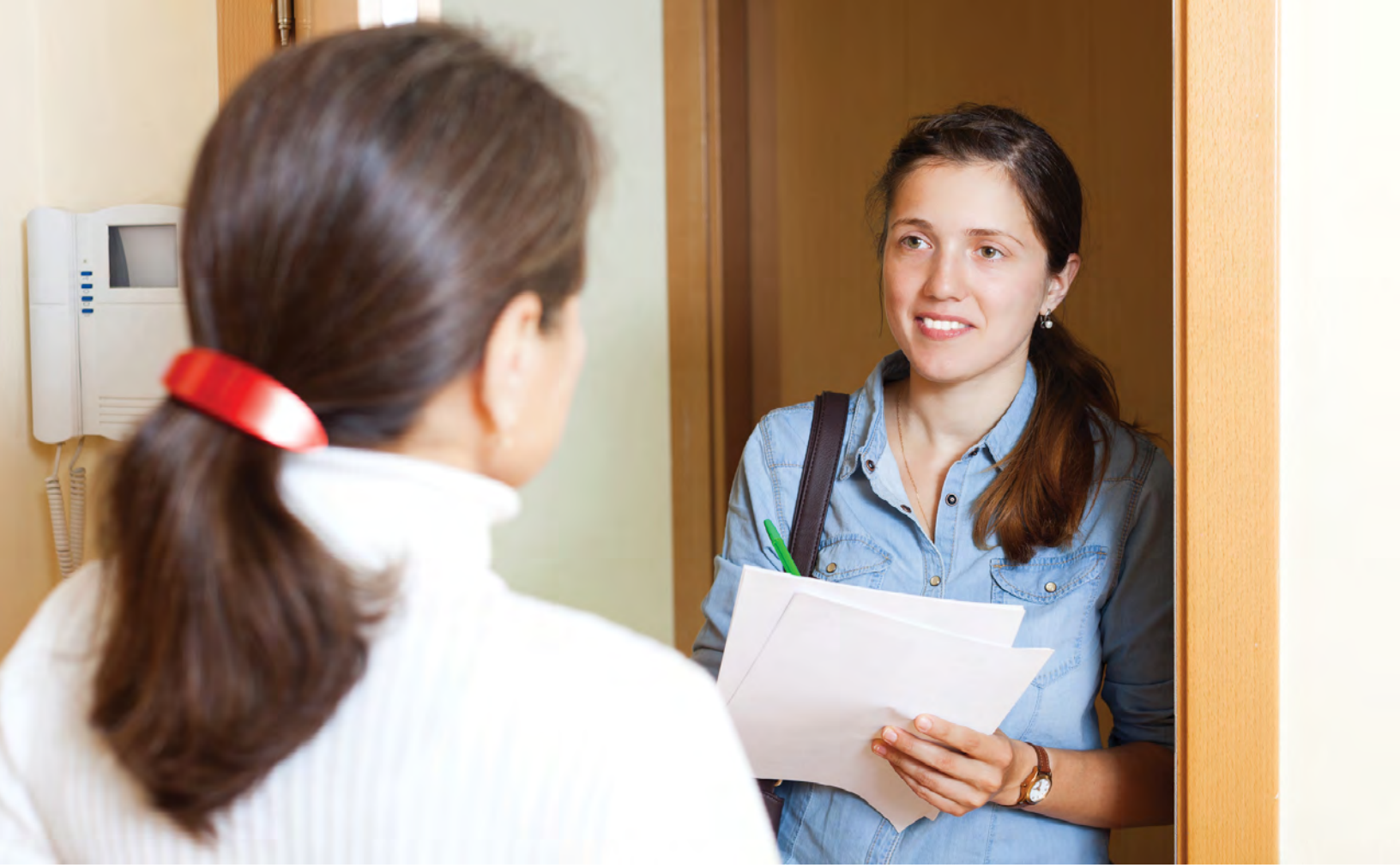


PHOTO: SHUTTERSTOCK.COM

LAYING THE GROUNDWORK FOR A RESEARCH PROJECT

LEVERAGING OUTSIDE OPPORTUNITIES

Launching a randomized evaluation requires a government to invest time and resources up front, while many of the evaluation's benefits are not realized until a much later date. We designed the J-PAL State and Local Innovation Initiative to help address this challenge. Through the initiative, we invite state and local governments to submit letters of interest describing how rigorous evidence would help answer an important policy question. We offer funding and technical support to selected governments to offset upfront costs and to help governments overcome barriers that might normally make research more difficult and expensive. This changes the cost-benefit analysis that governments often face when undertaking a research project.

More broadly, linking a research project to an external opportunity, such as a grant or competition, can be a useful strategy for building the support and momentum needed to get a project started. In addition to the J-PAL State and Local Innovation Initiative, there are a number of foundations and non-profits that offer opportunities that governments can leverage to build support for new research projects.[1]

COMMITMENTS FROM THE GOVERNMENT

While the J-PAL State and Local Innovation Initiative offers funding and technical support to offset some of the upfront costs of developing a randomized evaluation, we have found that the government partners who are successful in designing and launching randomized evaluations have also made a number of important commitments to the project.

In particular, successful research partnerships generally involve the commitment of (a) a senior-level decision-maker within the government, who ensures that the project aligns with the government's overall priorities, helps navigate relationships with key stakeholders, and provides momentum when needed, as well as (b) a day-to-day project manager, who allocates a significant percentage of his or her time to the project, serves as the point person for moving the project forward, and meets regularly with the researcher and other partners. J-PAL North America looks for evidence of these commitments when making awards through the State and Local Innovation Initiative.

Additionally, government partners bring valuable knowledge of the local context, including program operations, potential ethical or logistical constraints, and availability and quality of administrative data. In most cases, the government partner is also responsible for identifying funding to implement the program that will be evaluated. The government and research team can then work together to secure funding for any additional costs associated with the evaluation, applying either to J-PAL North America or to other funding sources.

WORKING WITH ACADEMIC RESEARCHERS

One of the goals of the State and Local Innovation Initiative is to connect state and local governments with academic researchers from J-PAL's network. Governments can benefit from researchers' technical expertise, and many researchers are interested in partnering with governments to study questions that are both of academic interest and relevant to government decision-makers.

In many cases, researchers receive university or private funding to take on an evaluation at no cost to the government partner. Other times, the researcher can work with government partners to raise funds for the evaluation from a foundation or other sources.

Some of the benefits of these kinds of policymaker and researcher partnerships include:

- *Technical expertise.* Researchers can bring expertise on a number of issues that would be difficult for individuals without training in evaluation methods to navigate on their own. This includes estimating the minimum sample size needed to detect a given change in outcomes, designing the randomized evaluation to minimize disruptions to service delivery, and identifying measures and data sources for outcomes of interest. Researchers may also have access to specialized research staff, such as survey designers, data analysts, or project managers.
- *Knowledge of the existing evidence.* Many researchers bring not only experience in evaluation methods, but also deep knowledge of the existing evidence on a particular topic. Researchers can support policymakers in interpreting this evidence and applying it to their own policies and programs.
- *Stability during leadership changes.* Partnering with an outside researcher can help sustain an evaluation after a transition in the administration or other leadership changes. Additionally, several of the governments that J-PAL North America has worked with have set up legal agreements between the government and research partner (such as Memoranda of Understanding or data sharing agreements), as a tool for encouraging future administrations to continue an evaluation.

Despite the many advantages, partnerships between governments and academic researchers remain relatively uncommon. In many cases, finding a research partner who has both an interest in the relevant topic and the time to take on a new evaluation can be challenging. J-PAL North America helps governments navigate this process by serving as a "matchmaker" to connect governments with potential research partners in our network.

[1] Harvard University's Faculty of Arts and Sciences produces a monthly "Big Data Funding Newsletter" featuring opportunities from government agencies and foundations around administrative data and other subjects related to data generation. This newsletter is available at bit.ly/2wR7iP1.

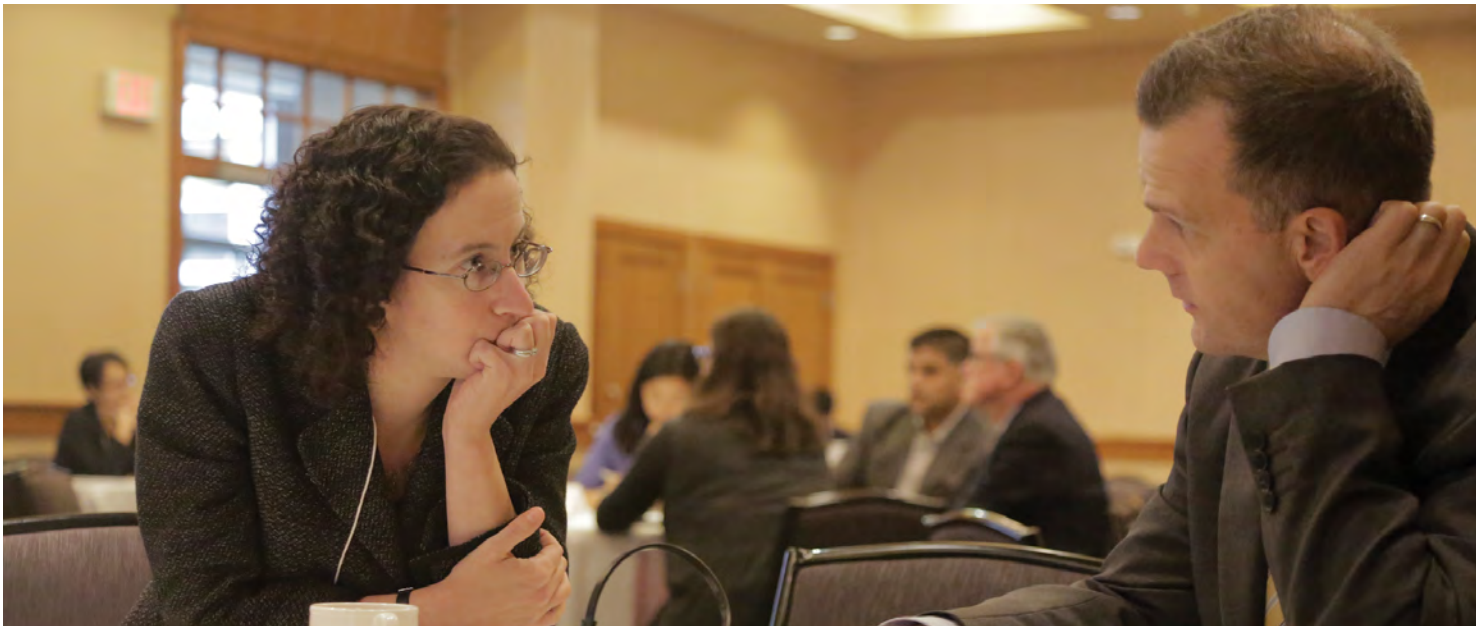


PHOTO: AMANDA KOHN | J-PAL/IPA

WHAT DOES A SUCCESSFUL RESEARCH PARTNERSHIP LOOK LIKE?

In our experience, a successful research partnership involves close collaboration between the researcher and the government to design a high-quality evaluation that is also politically, ethically, and logistically feasible.

In a successful research partnership, the government agency:

- Wants to better understand the impact of a policy or program
- Is implementing the policy or program at a sufficient scale, such that an evaluation will be able to detect meaningful changes in outcomes
- Is willing to think creatively about incorporating evaluation into program operations
- Facilitates access to administrative data

The researcher:

- Respects the agency's priorities and determines areas of substantive overlap with his or her own research interests
- Works with the government agency to assess the feasibility of an evaluation
- Is willing to think creatively about designing the evaluation to address practical, political and ethical concerns
- Helps the government navigate institutional or legal obstacles to sharing data

PHOTO: RKL _ FOTO | SHUTTERSTOCK.COM

IDENTIFYING OPPORTUNITIES FOR RANDOMIZED EVALUATION

DEFINING A RESEARCH QUESTION

In order to decide whether a randomized evaluation is the right method to use, it is important to start with a specific research question. Like other types of impact evaluations, the main purpose of randomized evaluations is to answer questions about a program's impact.

- *Impact questions* ask whether a program has an impact and how large that impact is. For example, "Does this tutoring program improve test scores?"

Impact evaluations can answer questions about introducing a new program or policy; implementing changes to an existing program or policy; or comparing two (or more) versions of a program or policy. Research questions can sometimes be reframed to be tested with an impact evaluation. For instance, a question like, "What type of staff should we be hiring for our tutoring program?" could be reframed as, "How does the impact of the tutoring program differ when delivered by volunteers, compared to when it is delivered by licensed teachers?"

Other kinds of research questions require different kinds of evaluations. For example:

- *Descriptive questions* ask about the nature and scope of a problem, including who is most affected. For example, "Which students in the school have the greatest need for tutoring services?" or "What challenges does this group of students face?" Needs assessments can often answer descriptive questions.
- *Process questions* ask how well the program is implemented. For example, "Are tutors showing up on time?" or "Do classrooms have the necessary supplies for tutoring?" Process evaluations can answer these types of questions.

Often, impact evaluations are conducted as part of a larger package of assessments. For example, it may be useful to begin by conducting a needs assessment in order to verify the existence or extent of a problem. A process evaluation can help understand whether a program is being delivered as intended and to the appropriate people. If an impact evaluation later finds that a program had no impact, a process evaluation can help distinguish implementation failure (a program was not delivered as planned) from an ineffective program (a program does not work even when delivered as planned).

Read more: Asking the right research question | bit.ly/2vUPfWV

DIFFERENT WAYS TO RANDOMIZE

Randomized evaluations require introducing some kind of random assignment into program or policy implementation. As described earlier, in the simplest form of a randomized evaluation, one begins by identifying the group of individuals who are eligible to participate in a program. These eligible individuals are then randomly assigned either to a treatment group, which can participate in the program, or to a control group, which cannot participate. There are also many variations on the simple form of a randomized evaluation. For example, instead of randomizing individual people into either the treatment or control group, it is possible to randomize larger units, such as a household, neighborhood, classroom, or school. Random assignment can be as simple as flipping a coin, rolling a die, or drawing names from a hat, but researchers typically randomize using statistical software.

For a randomized evaluation, random assignment must be proactively built into a program to determine who receives it. Only those individuals enrolled in a program through the random assignment process can be included in the evaluation. People who had already entered the program by other means cannot be included. However, a randomized evaluation does not necessarily require every program slot to be randomly assigned. For example, in an ongoing randomized evaluation of the WorkReady Philadelphia summer jobs program, the program ran a random lottery to select youth for 1,000 out of 8,000 program slots. The remaining 7,000 program slots were allocated at the discretion of the summer jobs providers.

WHEN DOES A RANDOMIZED EVALUATION MAKE SENSE?

While the feasibility of a particular randomized evaluation depends on a host of factors, below are some opportunities to look for:

- *Demand for a program exceeds the number of available program slots.* If limited resources prevent a program from serving everyone who is eligible, a lottery may be a fair alternative to allocating slots on a first come, first served basis. Lotteries can be particularly helpful when program administrators are interested in reaching people who might be less motivated to participate or who are not already connected to the service provider. In some cases, programs may have been planning to introduce additional eligibility criteria or other filtering mechanisms in order to reduce the number of eligible candidates to fit the available program slots. In these cases, random assignment can offer an alternative way of filtering that also enables a rigorous evaluation.
- *Rolling out or phasing in a program over time.* A program that will eventually serve every individual or unit in the target area might be difficult to launch everywhere at once for operational reasons. Rather than rolling out the program in an ad hoc manner over time, the order in which units receive the program can be randomized. The individuals or units who receive the program later serve as a control group to compare to those who receive it earlier.
- *Adding a new intervention to an existing program.* Randomizing program participants to receive different versions of the program creates an opportunity to test the impact of the new intervention relative to the original program. This evaluation design can also test the impact of each program relative to a pure control group, which does not receive any version of the program.

- *Refining or reconsidering program eligibility criteria.* When individuals or groups are scored on some eligibility criteria, it is possible to randomize those people whose scores are just within or outside the eligibility cutoff (i.e., "on the bubble") into or out of the program. Meanwhile, people well within the program eligibility cutoff would automatically receive the program, and those well outside the cutoff would not qualify. This evaluation design can help determine whether the program is effective for people just outside the eligibility cutoff and whether eligibility should be expanded.
- *An entitlement program has low take-up.* Individuals who are eligible but not yet participating in the program can be randomly assigned to receive encouragement to enroll, such as by letters in the mail, phone calls, or text messages. In this instance, the randomized evaluation can help answer the question of how to effectively encourage more people to participate in the program. Additionally, if the sample size is large enough and the encouragement has a big effect on participation, researchers can evaluate the impact of the program itself by comparing those who received the encouragement to those who did not. This enables rigorous evaluation of a program, without denying anyone access to the program.

EXAMPLE: THE PROGRESA EVALUATION

In 1998, the Mexican government pursued a phase-in randomized evaluation to study the impact of a national conditional cash transfer program called PROGRESA. The government could not launch PROGRESA in all 506 eligible villages at the same time due to administrative and budget constraints, so 320 villages were assigned to the treatment group to receive cash transfers immediately and 185 villages were assigned to the comparison group and received cash transfers two years later.

Read more about the PROGRESA evaluation on health outcomes and school enrollment:

Health outcomes | bit.ly/2imXtCB

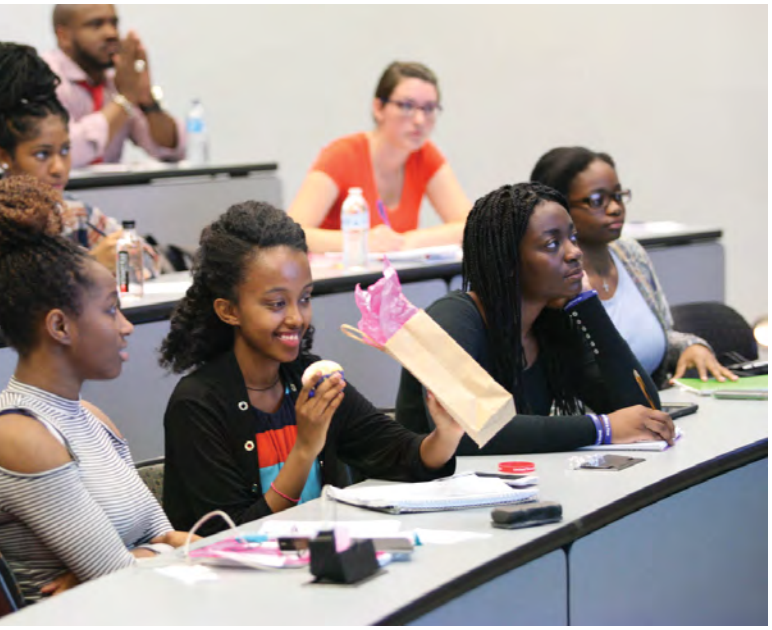
School participation | bit.ly/2wuqeTV

EXAMPLE: THE SNAP TAKE-UP EVALUATION

Although the Supplemental Nutrition Assistance Program (SNAP) is an entitlement program, only 41 percent of eligible elderly individuals had enrolled in 2013. The nonprofit organization Benefits Data Trust (BDT) provides targeted outreach and comprehensive application assistance to individuals who are likely eligible for SNAP and other programs. BDT has partnered with J-PAL affiliates Amy Finkelstein (MIT) and Matthew Notowidigdo (Northwestern University) to conduct an ongoing randomized evaluation of the effect of informational mailings and application assistance on SNAP enrollment in Pennsylvania. The evaluation will examine the effect on SNAP enrollment of two different interventions—a low-intensity informational mailing and high-intensity outreach with SNAP application assistance. Knowledge generated by the evaluation will help BDT understand which outreach activities are most effective at connecting eligible households to SNAP and better target its efforts in the future.

Read more: The SNAP take-up evaluation | bit.ly/2wkpdgK

CASE STUDY: PHILADELPHIA



A WORKREADY PHILADELPHIA PARTICIPANT HOLDS ONE OF THE PRODUCTS DESIGNED BY YOUTH IN THE PROGRAM. PHOTO: PHILADELPHIA YOUTH NETWORK

How can a randomized evaluation be designed to minimize disruption to the usual recruitment and enrollment processes?

Randomized evaluations are not one-size-fits-all; rather, they can be thoughtfully tailored to minimize disruption for programs implemented by multiple service providers or that involve multiple service models. The ongoing randomized evaluation of Philadelphia's WorkReady summer jobs program involves both of these factors. The research team has worked collaboratively with the City of Philadelphia and the service providers to design a study that will provide rigorous evidence on the impact of the program while minimizing any disruption to program operations.

Led by J-PAL affiliate Sara Heller (University of Michigan), the randomized evaluation will test the impact of being offered a summer job through Philadelphia's WorkReady program on criminal justice, employment, and education outcomes. Young people in Philadelphia face challenges common to youth in low-income neighborhoods across the United States—high rates of dropout, lack of employment opportunities, and exposure to violent crime. Previous randomized evaluations in New York City and Chicago found that summer jobs programs led to a drop in violent crime, incarceration, and even mortality.[2] The evaluation in Philadelphia will test whether these results apply in a new setting, as well as whether the summer jobs program impacts other outcomes such as mental health, substance abuse, teen pregnancy, housing instability, and child maltreatment.

WorkReady Philadelphia is a portfolio of programs that address the skills gap for vulnerable young people, managed by the Philadelphia Youth Network (PYN). PYN has offered youth employment training skills and work experience through the WorkReady summer jobs program for 15 years. Sixty local agencies contract with PYN to place youth in six-week (120-hour) summer jobs. Given limited resources, demand for the program consistently outpaces available positions. In 2016, for example, 16,000 youth applied for approximately 8,000 summer jobs. In the past, jobs were awarded by provider discretion, screening processes, or on a first-come, first-served basis. Applicants were matched to jobs based on geographic proximity and experience. This approach meant that many youth who received summer jobs perhaps may have been more likely to find summer opportunities without WorkReady. On the other hand, youth who were less likely to be selected by providers may have actually been those who would benefit most from the program. Analysis of program data showed, for example, that young people of color were less likely to be placed in jobs.

Committing to a more equitable distribution of program slots in 2017, PYN agreed to randomly allocate roughly 1,000 of its 8,000 program slots by a fair lottery—which would also enable a rigorous evaluation of the program. The remaining 7,000 program slots would be allocated as usual. Only youth whose participation was determined by lottery will be included in the evaluation.

WorkReady providers and the research team placed paramount importance on implementing the lottery in a way that placed youth in appropriate jobs while retaining random assignment. Youth who received jobs would need a reasonable commute to their workplace, so assigning individuals to difficult-to-reach positions could not only create obstacles for the youth and the providers, but also negatively impact the research—far-flung job placements could lower compliance with the program (i.e., increase dropout), reducing the researchers' ability to estimate the impact of the program.

To address this potential issue, researchers designed a randomization strategy that included geographic blocking based on the preferences of each provider. Applicants were subdivided into pools by the geographic catchment area appropriate for specific jobs and then randomized to either the treatment or control group for those positions.

The research design also accounted for the fact that not every summer job is appropriate for every applicant. The WorkReady program offers four program models to meet the needs of different populations and a range of ages from 12-21. Three of these models were included in the study: service learning for youth with little or no prior work experience, structured work experience for youth with little or no prior experience, and internships for youth already prepared for the workplace. To accommodate the multiple service models, eligible youth were first categorized based on age (in addition to geographic area). Youth were then randomized within these age categories into either the treatment group or control group, so that they were only assigned to job models appropriate for their age.

PYN added new recruitment and enrollment supports as part of the evaluation, which also aligned with the City's goal of enrolling more disadvantaged youth in the program. To make the randomized evaluation informative, it was important to ensure that take-up rates in the treatment group were high. If few youth in the treatment group accepted and completed their summer job, the effects of the program on participating youth would be diluted by individuals in the treatment group who had not actually received the intervention, and the randomized evaluation would underestimate the program's impact.

To encourage the high take-up needed for an informative evaluation, PYN hired a recruitment specialist and additional support staff. One barrier to high take-up was potentially burdensome paperwork requirements for accepting the job. For individuals assigned to the treatment group, the recruitment specialist and support staff encouraged them to accept the job and followed up to make sure they completed the required paperwork. The WorkReady program was already seeking to expand to and engage youth facing barriers to employment such as criminal justice involvement, pregnancy or parenting responsibilities, and unstable housing. Increasing take-up support helped the City, PYN, and service providers achieve their goals by assisting youth who might not otherwise enroll in or complete the program.

[2] Read about the New York City evaluation at bit.ly/2fcQ2K8 and the Chicago evaluation at bit.ly/2f3IRDE.

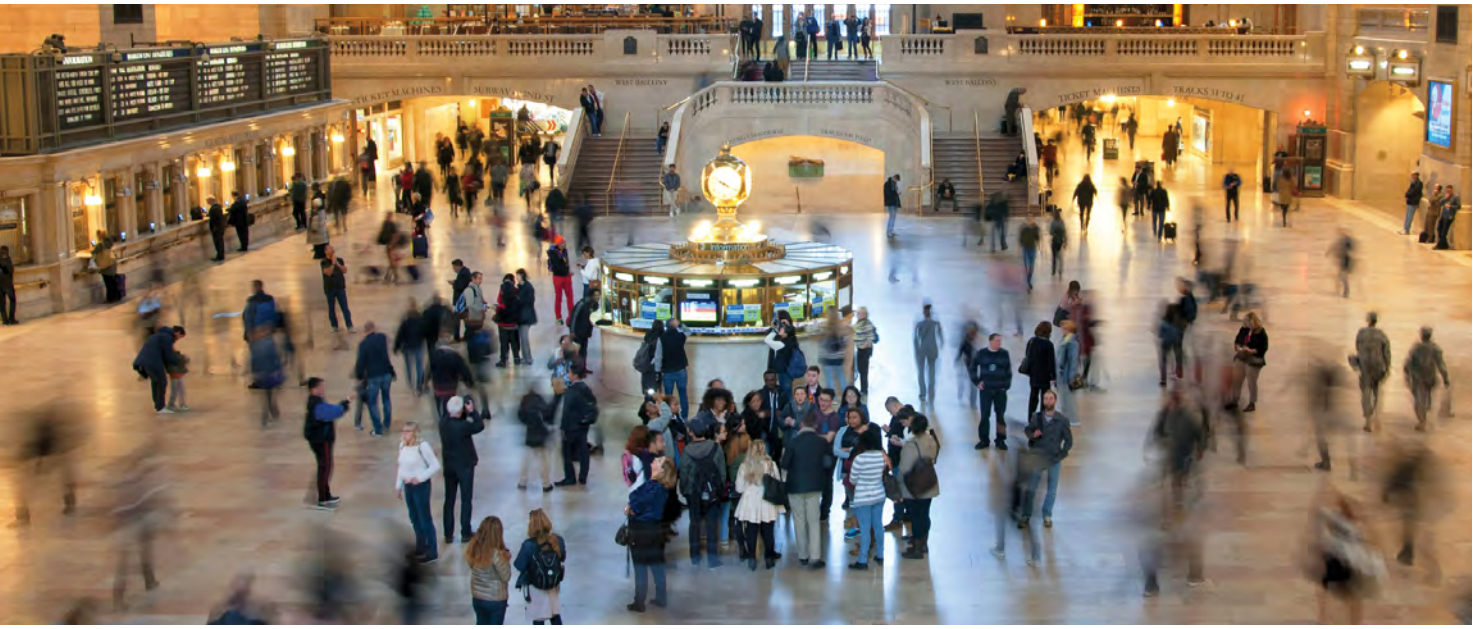


PHOTO: CATE _ 89 | SHUTTERSTOCK.COM

HOW LARGE A SAMPLE SIZE WOULD AN EVALUATION NEED?

One of the most frequent questions we get from government agencies or other partners that are considering an evaluation is about sample size, or how many people will be randomly assigned to the treatment and control groups as part of the evaluation. There is no simple answer to this question, because the optimal sample size for an evaluation depends on many factors. In general, larger sample sizes enable researchers to detect even small differences in outcomes between the treatment and control groups, whereas smaller sample sizes can only detect large differences in outcomes. For this reason, randomized evaluations typically have sample sizes ranging from the hundreds to the thousands.

The statistical power of an evaluation reflects how likely we are to detect any meaningful changes in an outcome of interest brought about by a successful program. When determining the sample size needed for an evaluation, it can be helpful to consider the smallest effect that would make a program worthwhile. For example, would it be worth continuing an employment program that increased participants' annual earnings by only $100? Or, would the program only be worthwhile if it increased participants' earnings by $1,000? Researchers can then run "power calculations" to estimate the sample size that would be needed to detect a given effect size.

Randomized evaluations that use a clustered design (i.e., that randomize groups rather than individuals) generally require a larger sample size. For example, a randomized evaluation that assigned individual students to treatment or control would have greater statistical power than a randomized evaluation that assigned classrooms to treatment or control, even if the total number of students participating in the evaluation was the same in both cases.

Given the great amount of time and resources that often go into evaluation, we often recommend against running evaluations where the sample size appears to be too small to detect meaningful effects.

Read more: The risk of underpowered evaluations | bit.ly/2vlEjDc

WHEN NOT TO DO A RANDOMIZED EVALUATION

There are a number of circumstances in which a randomized evaluation would not be feasible or appropriate, including when:

- *There is strong evidence that the program has a positive impact and we have the resources to serve everyone who is eligible.* It would be unethical to deny people access to a program that has been proven to be effective for no reason other than conducting a randomized evaluation. Under these conditions, resources would be better spent ensuring that the program continues to be implemented as intended and/or scaling up the program so that more people can benefit.
- *The program's implementation is changing.* Evaluating a program while the implementation is changing could yield results that would be difficult to interpret. For example, suppose that a tutoring program shifted from being mandatory during the day to optional and after school midway through an evaluation. The results of the evaluation would represent the average impact of both approaches. If the evaluation found a positive impact, it could be because both approaches had a positive impact, or because one approach had a positive impact and the other had no impact or even a negative impact.
- *The sample size is too small.* If researchers believe that the potential sample size is too small to be able to detect meaningful changes in outcomes, then there is a risk that the evaluation could consume time and financial resources but produce only inconclusive results. Imagine, for example, a randomized evaluation of a tutoring program that found that the program increased test scores by 10 percent, but that increase was not statistically significant. We would not be sure whether the program had a positive impact or whether the increase was due to chance.
- *The time and financial cost outweigh the potential benefits of the evidence generated.* Governments should always weigh the potential costs of an evaluation against the value of the evidence generated. In some cases, answering a particular question will require a large investment of time or other resources (for example, because the outcomes of interest are difficult to measure or can only be measured after significant time has passed). If the evaluation would answer a question of great importance to the government or others, then it may still be worth pursuing. If the evaluation is unlikely to provide new insights or influence decision-making, then those resources may be better spent elsewhere.

CASE STUDY: PENNSYLVANIA

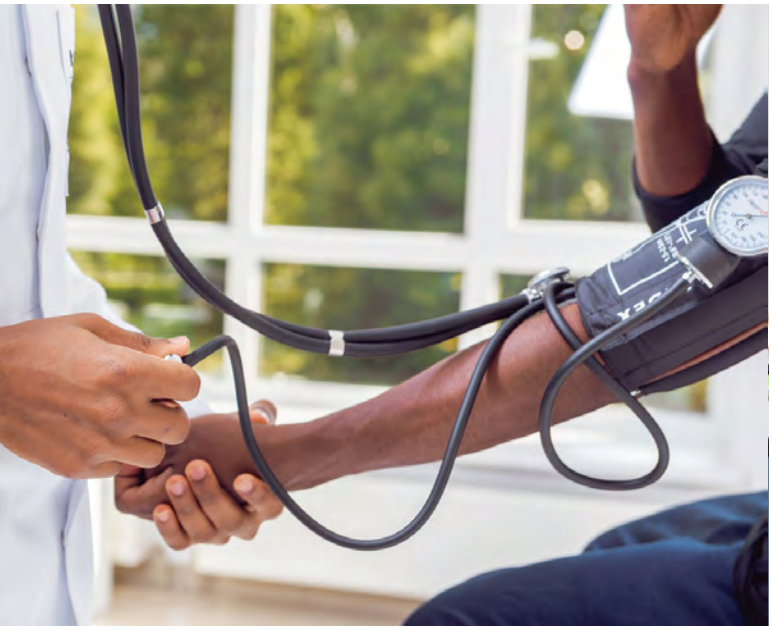


PHOTO: SHUTTERSTOCK.COM

What can be learned when the intended evaluation turns out not to be feasible?

There are several reasons why, after beginning to pursue a randomized evaluation, the government or research team may decide not to proceed. For example, resources may become available to allow the program to serve everyone who is eligible, which may make it impractical or unethical to randomly assign eligible individuals to a control group. In other cases, the government or research team may decide that they need to refine how the program is implemented before undertaking an evaluation.

However, even when a randomized evaluation is not launched, the process of developing the study can still provide useful intermediate outputs that can help the government achieve other research and policy goals.

In 2016, the Commonwealth of Pennsylvania announced funding for 45 Centers of Excellence (COEs), which are designed to coordinate care for individuals with opioid use disorder to help ensure that they stay in treatment, receive follow-up care, and are supported within their communities. The COEs deploy care management teams to assess patient needs and develop a treatment plan, make warm hand-offs to physical health, mental health, and substance use treatment providers, and issue referrals for employment, housing, and legal services. Staff from the Governor's Office and the Department of Human Services partnered with J-PAL North America through the State and Local Innovation Initiative to explore using a randomized evaluation to better understand whether COEs effectively increase engagement with treatment and, if so, which components of the model are most effective.

After working with J-PAL North America staff and researchers, Pennsylvania ultimately decided that a randomized evaluation of the COEs would not be feasible at this time. The proposed evaluation would be implemented across a number of different COEs, and discussions with staff from various COEs revealed wide variation in care coordination practices, including variation in which staff deliver services and in what additional services the COEs provide. Because a randomized evaluation would estimate the average effect across different COEs, this variation would make it difficult to interpret the results of a randomized evaluation.

For example, suppose that some COEs deployed peer counselors to coach participants in a community setting while other COEs hired nurses to support participants in a more clinical setting. A randomized evaluation that found no impact of participating in a COE could imply that that neither model was effective. It could also imply that one model was effective but the other was ineffective, so that the overall impact was, on average, insignificant. Conversely, if a randomized evaluation found positive impacts from participating in a COE, we would want to know which model had produced the result, so it could potentially be replicated elsewhere.

Even though a randomized evaluation was not launched, the initial work staff did to develop a randomized evaluation was useful in thinking about how to measure the impact of the state's many efforts to address the opioid and heroin epidemic. For example, in the process of scoping a randomized evaluation of the COEs, staff from Pennsylvania discussed how to measure outcomes such as persistence in treatment and health care utilization. The metrics and potential data sources they identified have been used for other grants and projects, including Pennsylvania's successful funding application for the 21st Century CURES Act, and can also serve as a starting point as the state considers future opportunities for evaluation.

By partnering with J-PAL North America and academic researchers to pursue a randomized evaluation, Pennsylvania was able to access both external financial support and technical expertise, which has helped to uncover other potential opportunities for evaluation. For example, in a meeting with staff representing several different COEs, a service provider observed that demand often exceeds capacity for detox beds, leading to limited and intermittent detox bed availability. A researcher from J-PAL's network who also participated in the discussion noted the possibility of conducting a quasi-experimental evaluation to measure the impact of detox bed availability. J-PAL North America staff connected the researcher with Pennsylvania's Department of Human Sevices to continue the conversation.

DEFINING AND MEASURING OUTCOMES

Usually, a research question will identify the primary outcomes of interest in an evaluation. As a government agency moves forward in developing a randomized evaluation, it will also need to identify which indicators will be used to measure that outcome and how to collect data on those indicators.

Consider, for example the Moving to Opportunity project, which the US Department of Housing and Urban Development (HUD) launched in 1994 to test the impact of offering housing vouchers to families living in high-poverty neighborhoods. HUD and its research partners were interested in measuring the impact of the housing voucher programs on health, but they first had to identify specific indicators that could be used to measure health. The researchers conducted surveys asking individuals about their physical and mental health. Survey personnel also visited individuals to measure obesity (by recording their height and weight) and likelihood of diabetes (using blood glucose tests).

Below are a few considerations to keep in mind when choosing outcomes, indicators, and data sources for a randomized evaluation.

Where does data currently exist?

There are generally two kinds of data that can be used to measure outcomes: primary data, which researchers collect themselves as part of an evaluation, and secondary data, which is available from other sources. Administrative data—data collected routinely by the government or other organizations for non-research purposes—are a type of secondary data. For instance, school systems routinely track grades, attendance and graduation for students.

Benefits of using administrative data for research include:

- *Lower cost and greater ease.* Using administrative data eliminates the need to develop surveys, hire a survey firm, or track down the individuals who are part of the evaluation.
- *Reduced burden on participants.* Individuals are not asked to share information that has already been collected elsewhere.
- *Near-universal coverage of the individuals in the evaluation.* Unlike surveys, individuals do not need to respond actively to be covered in administrative data.
- *Greater accuracy and lower risk of bias.* Administrative data may be more accurate than surveys in measuring outcomes that are sensitive or hard to remember.
- *Long-term availability.* Administrative data are collected regularly over time, enabling researchers to measure long-term outcomes without needing to track down individuals years later.

Because of these benefits, it can be helpful to think about potential sources of administrative data when selecting outcomes and indicators. Potential sources of administrative data include not only data that the government running a program collects, but also data collected by other levels of government or other organizations that could be used for research purposes.

Of course, it may be important for an evaluation to measure outcomes that are not available in administrative data, and some evaluations use both administrative and survey data. For example, the Moving to Opportunity project used survey data to track the health outcomes of adults who moved after receiving housing vouchers. While it was both expensive and time-consuming for surveyors to collect original data to measure obesity and likelihood of diabetes, doing so allowed researchers to confirm that moving to a low poverty neighborhood caused people to be healthier across both dimensions. Researchers also used administrative data to examine long-term outcomes for the children in families who moved to low-poverty neighborhoods. Administrative data from IRS tax records enabled a nearly 20-year follow up study demonstrating that young children who moved to low-poverty neighborhoods had higher rates of college attendance, earned higher incomes, were less likely to become single parents, and were themselves more likely to live in better neighborhoods.

Read more:
The Moving to Opportunity project | bit.ly/2cpApxS
Resources on administrative data | bit.ly/2wk8eeq

Are data available for members of both the treatment and control groups?

For a randomized evaluation, data must be available for both members of the treatment group, who participate in the program, and members of the control group, who do not. Therefore, data collected by the program being evaluated (such as information provided on a program's intake form or surveys conducted by program staff) generally cannot be used to measure outcomes.

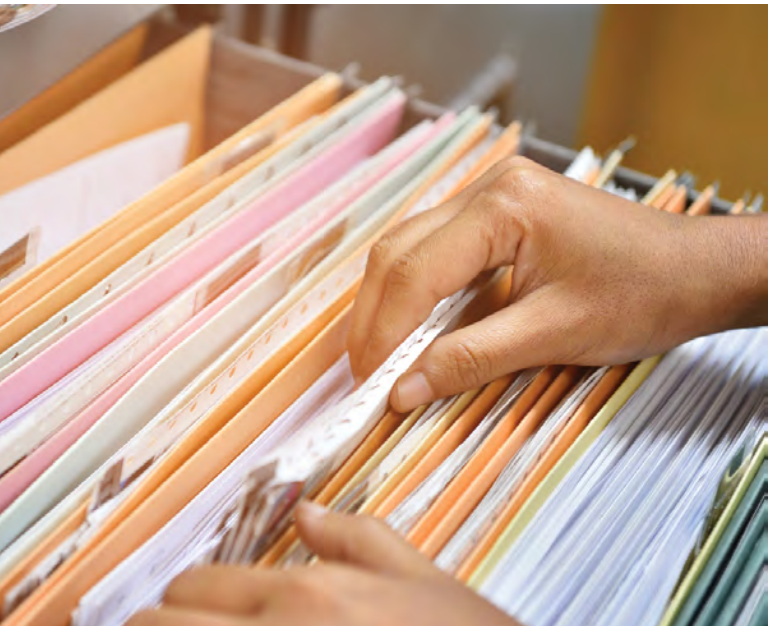


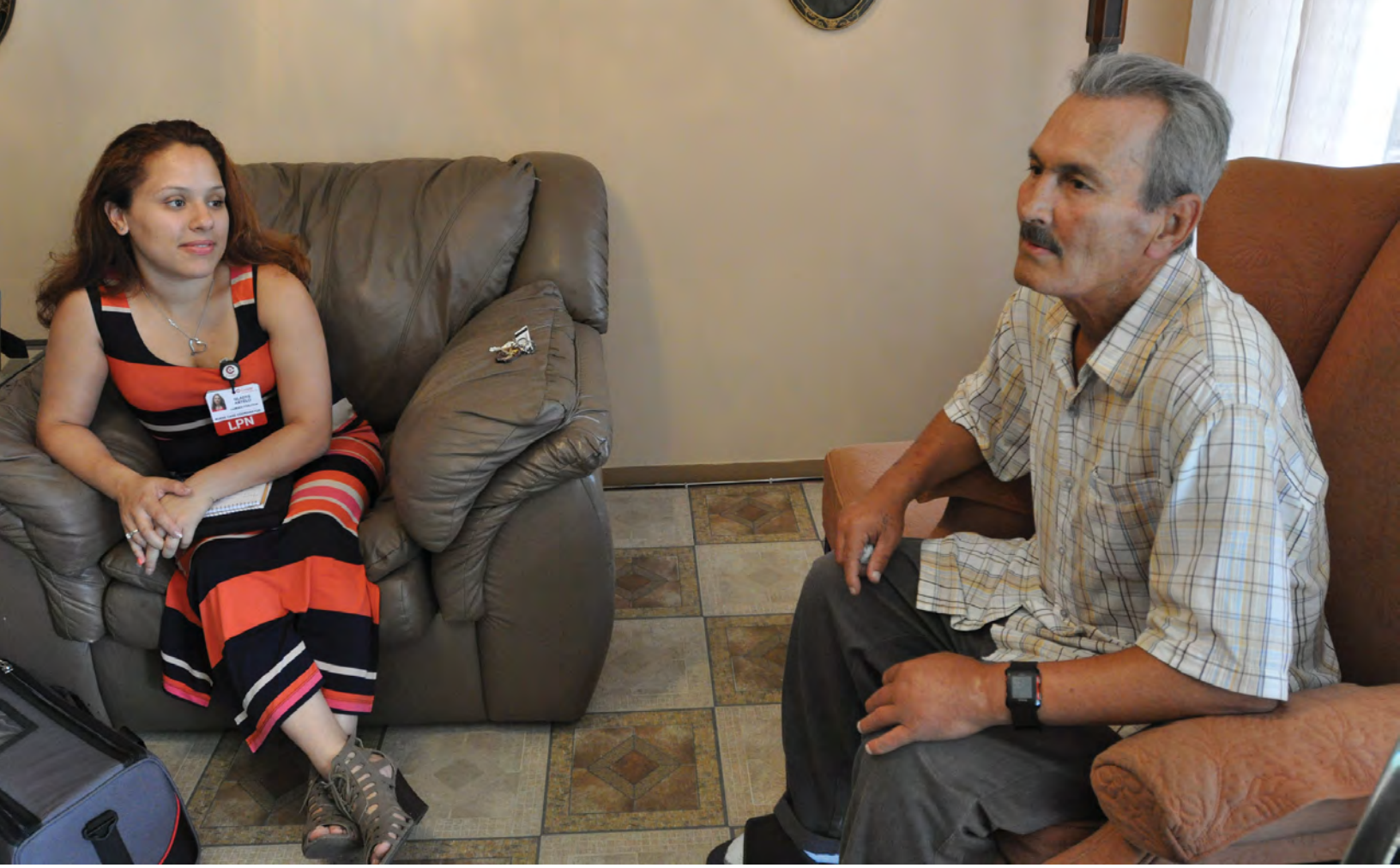
PHOTO: SHUTTERSTOCK.COM

Additionally, researchers must be able to match the outcomes data to records of whether an individual is in the treatment group or the control group. If random assignment is carried out at an individual level, it will be necessary to measure outcomes using individual-level data (as opposed to data that is only available at the school or neighborhood level, for example).

How removed are the outcomes of interest from the program or intervention?

Before beginning a randomized evaluation, it can be useful to lay out a theory of change or logic model that describes the pathway through which the program expects to achieve its desired impact, and to identify intermediate outcomes that can measure each step along this theory of change. That way, if a randomized evaluation finds that the program did not have the impact that was expected, it will be easier to identify which steps in the program's theory of change were not correct.

Additionally, if the program can only affect the key outcomes of interest through many steps, the magnitude of any reasonable change may be so small that it would be difficult to detect without a very large sample size. Before beginning an evaluation, it is important to consider what kind of change in the outcome of interest is reasonable to expect, and to discuss with researchers whether the planned study would have sufficient statistical power to detect changes of that magnitude.



A NURSE FROM THE CAMDEN COALITION OF HEALTHCARE PROVIDERS WORKS WITH A CLIENT. PHOTO: JOHN TEBES | J-PAL/IPA

IMPLEMENTING AN EVALUATION

ETHICS AND INSTITUTIONAL REVIEW BOARDS

Any research involving human subjects must be carried out in accordance with the principles of ethical research. In the United States, ethical principles and guidelines for research involving human subjects are laid out in the Belmont Report. These basic ethical principles are: (1) respect for persons, (2) beneficence, and (3) justice.

Most academic researchers have Institutional Review Boards (IRBs) at their host university that will review any research involving human subjects. Some government agencies may also have their own IRBs. Before research begins, an IRB must review and approve research protocols, such as procedures for obtaining informed consent, any surveys or questionnaires, and plans for how data will be shared and stored.

Read more:
The Belmont Report | bit.ly/2uvrnZO
The South Carolina Nurse Family Partnership pay for success pilot | bit.ly/2fLXCOA

WORKING WITH SERVICE PROVIDERS AND OTHER STAKEHOLDERS

Except in cases where a program was already using a lottery or some other form of random assignment to select participants, implementing a randomized evaluation will require at least some changes to how a program operates. Researchers will design and implement the randomized evaluation with the goal of limiting the number of these changes. For those that cannot be avoided, it is important to consider how they may affect stakeholders and identify ways to mitigate the impact of the changes.

A few common areas of sensitivity include:

- *Will program staff interact with individuals who are randomly assigned to the control group?* In some evaluations, research staff are responsible for screening potential participants for eligibility, obtaining informed consent, and randomly assigning individuals to either the treatment or control group. In other cases, program staff will carry out these responsibilities. While this has some advantages (for example, program staff may be more familiar with how to explain the program to potential participants), it can also be difficult for program staff to tell individuals who are randomized to the control group that they will not be able to participate in the program. Frequent communication with program staff to explain the goals of the randomized evaluation can help build their support. Additionally, the research team and the service provider can offer training and ongoing support to program staff in navigating these difficult conversations.
- *Is there a chance that there will be people assigned to a control group while program slots go unfilled?* Particularly if a program is new or expanding, it may take some time for the program to fill all of its slots, which can create pressure to enroll people who were originally assigned to the control group. There are a variety of strategies that can be used to ensure that all slots are filled without compromising the randomized evaluation, such as maintaining a randomly ordered waitlist or temporarily increasing the percentage of people who are assigned to the treatment group. Planning in advance which strategies a program will use to address this contingency can help ease anxieties.
- *Will new data requests place a burden on data providers?* Randomized evaluations that use administrative data require close collaboration with staff from the data provider (which may or may not be the same organization that is delivering the program begin evaluated). In exchange for their help providing data for the evaluation, researchers may be able to provide pro bono support with other analytic tasks. In some cases, researchers may even provide an intern or research assistant, who can work directly with the data provider to help carry out the data requests.

THE IMPORTANCE OF PILOTING

Before launching a randomized evaluation, it can be useful to build in a pilot period to ensure that the program is being delivered as intended and that the research and program staff can successfully carry out any new protocols required for the randomized evaluation (e.g., obtaining informed consent, administering baseline surveys, randomizing people into treatment and control). Pilots can be especially useful when the program being evaluated is new and still working out logistics, or in cases when the randomized evaluation requires significant changes to program operations. Below are some examples of questions that a pilot can help answer:

- *Recruitment.* Is the program able to recruit a sufficient number of individuals to both fill the available slots and create a control group?
- *Take-up.* What percentage of people assigned to the treatment group actually enroll in the program? A low take-up rate can negatively affect the statistical power of an evaluation.
- *Crossovers.* Are people who were originally assigned to the control group participating in the program? A large number of crossovers can make it difficult to estimate the impact of the program. Imagine, for example, that everyone in the control group actually participated in the program. In that instance, we could no longer use the control group to estimate what would have happened to participants in the absence of the program.
- *Data sharing.* Are systems in place to share data securely? Are there problems of missing data or poor data quality?

Depending on the circumstances of the pilot, researchers may include outcomes for people who enrolled during the pilot period in the evaluation results. In some cases, the government and the research team may determine, after a pilot period, that it does not make sense to proceed with a full evaluation.

Read more: The importance of piloting an RCT intervention | bit.ly/2xrX05z

CASE STUDY: ROCHESTER

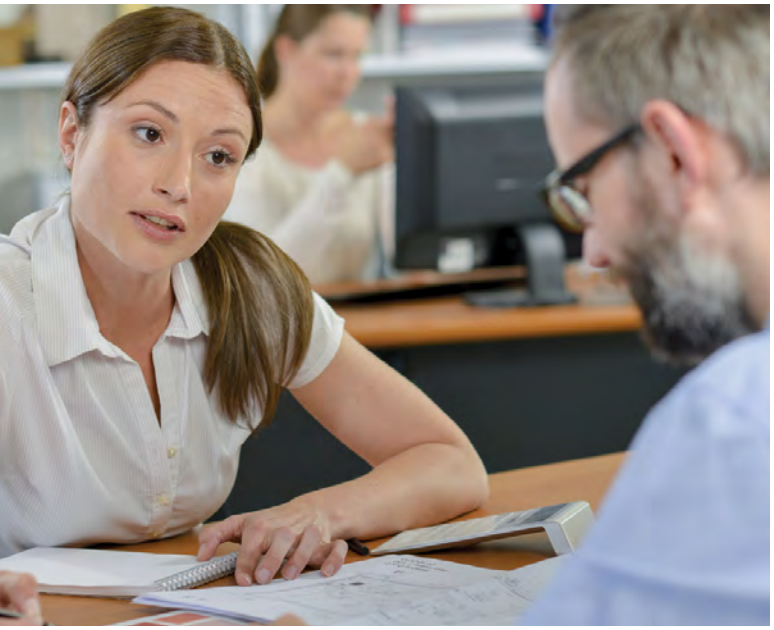


PHOTO: SHUTTERSTOCK.COM

How can a randomized evaluation be designed to address service providers' concerns?

When a program lacks resources to serve everyone who is eligible, random assignment can be a fair way to allocate limited slots and a rigorous way to evaluate the program's impact. With a new program, however, it can be difficult to predict whether or not there will be more eligible people than slots available. Likewise, with social service programs that use rolling enrollment, it is not feasible to do the entire recruitment at once to know with certainty that the program has more eligible applicants than slots available. The evaluation of the Bridges to Success program in Rochester illustrates how service providers and researchers can navigate uncertainty around program enrollment and ensure that the evaluation does not reduce the number of people who would have otherwise received services.

The Rochester-Monroe Anti-Poverty Initiative (RMAPI), the City of Rochester, Action for a Better Community, the Catholic Family Center, and the Community Place of Greater Rochester are currently piloting Bridges to Success, an innovative adult mentor/navigator program that aims to help residents in high-poverty neighborhoods of Rochester overcome barriers to self sufficiency. Professional navigators will help program participants set and achieve specific goals related to family and financial stability, health, and employment through coaching and referrals to an established network of service providers. At the same time, employment and dependent liaisons will support career-readiness and effective parenting, respectively.

To measure the impact of Bridges to Success, City officials and service providers in Rochester partnered with the Wilson Sheehan Lab for Economic Opportunities (LEO) at the University of Notre Dame and J-PAL North America to conduct a randomized evaluation. Researchers Bill Evans (University of Notre Dame), Javier Espinosa (Rochester Institute of Technology), and David Phillips (University of Notre Dame) are leading the evaluation.

The service providers expected that the number of people who would be eligible for the program would be greater than the number of people who could be served during the pilot period. However, they also recognized that other social services had been underutilized in the past and were wary that the program may not enroll enough individuals. If too few people enrolled, assigning individuals to a control group could deny residents access to a potentially beneficial program that still had available slots. Meanwhile, over-recruiting could create the feeling that social service providers were drumming up interest from the community only to let down the people they could not serve. The researchers stressed that they did not want to deny services if there was room to serve more people in the Bridges to Success program. To ensure that everyone would receive some form of assistance even if they were not assigned to Bridges to Success, individuals in the control group would receive a warm hand-off to existing social service providers.

The research team and service providers agreed that the study should not compromise service delivery and developed a contingency plan. In order to test whether the program would be underutilized, the service providers would pilot the program with a small fraction of the study sample before launching a full evaluation. Judging by the rate of intake to the program, the size of the target population, and the number of spots remaining in the program, the research team would be able to determine whether or not there would be enough demand to fill the program slots and have a control group. If enrollment trends in the pilot suggested that there would not be enough enrolled participants, the program team could reexamine eligibility and geographical target areas to ensure that services would not be denied

due to the research. In the unlikely event that not all of the program slots could be filled, the research team had a further contingency of randomly offering the program to people in the control group to fill the remaining program slots. These contingencies were designed to ensure that the evaluation would not proceed if having a control group would require leaving program slots unfilled.

To ensure that enough people participated while still serving the appropriate residents, J-PAL North America, LEO, the City of Rochester, and the service providers considered how they might expand the number of eligible individuals. In initial designs, applicants to the program needed to have annual wages of no more than 175 percent of the federal poverty level, have a stated desire to maintain full-time employment, be a head of household able to work (i.e., not receiving disability benefits), and have a high school diploma or GED. However, service providers identified many people who seemed well positioned to succeed in the Bridges to Success program despite lacking a diploma or GED. The research team and service providers adjusted the eligibility criteria so that it was not necessary to have a high school diploma or GED in order to participate. Removing this eligibility criteria would allow the program to serve more individuals likely to benefit without compromising the program's targeting.

One potential concern about denial of service related to the need for Bridges to Success to operate distinctly from a concurrent intervention implemented by the Catholic Family Center called the Family Independence Initiative. Like Bridges to Success, the Family Independence Initiative aims to help individuals achieve self-sufficiency, but it uses peer networks rather than mentor/navigators. If individuals assigned to the control group in the Bridges to Success evaluation participated in the Family Independence Initiative at higher rates than individuals assigned to the treatment group, it might be difficult to interpret the results of the evaluation. Imagine, for example, that everyone in the control group enrolled in the Family Independence Initiative. In that scenario, the evaluation would capture the relative effectiveness of the two programs, rather than the effectiveness of Bridges to Success relative to the status quo. In response to these concerns, participation in each program would be closely monitored to avoid overlap that could muddle the results of the evaluation.

Clear communication through weekly calls helped the research team and multiple service providers to develop a contingency plan, reconsider eligibility criteria, and plan to implement two different interventions separately. Additionally, having a research partner on the ground (Javier Espinosa, Rochester Institute of Technology) helped the research team stay abreast of progress and participate in meetings with service providers. The Catholic Family Center acted as the key point of contact to coordinate with the service providers. Clear and frequent communication helped solidify trust between all the parties involved, building on the strong coalition of agencies and the City of Rochester and clarifying shared priorities across the research team and implementing partners. With program enrollment underway and consensus on how to address challenges that arise during implementation, this coalition looks forward to increasing the value of the study by working to link administrative data to calculate return on investment.

MANAGING EXPECTATIONS AROUND COMMUNICATION

Early on in a research partnership, it can be helpful to establish clear timelines around when results will be available and when and how those results will be communicated. Some examples of expectations to discuss upfront include:

- When researchers expect to publish the results of the study.
- Whether the researchers will be able to make preliminary results available to the government partner and, if so, what they expect will be learned from those preliminary results.
- Whether the researchers may consider extending the timeframe of the evaluation, in order to increase the number of individuals enrolled in the study or to collect more outcomes data.

Almost all academic researchers will insist on having the results of the study be made public, whether they show a positive, a negative, or no effect. Researchers may also be working with other organizations that have strong preferences or requirements on how results are shared. For example, some academic journals insist that any studies published in the journal be kept confidential until publication.

Making a pre-analysis plan is one tool that can help manage expectations around how results will be communicated. A pre-analysis plan specifies how the researchers will analyze the data, including which hypotheses they will test and which data sources they will use to measure outcomes. A pre-analysis plan can also specify in advance how many individuals will be enrolled in the study and over what timeframe, how and when studies might be extended or turned into follow-up studies, or how results will be shared externally.

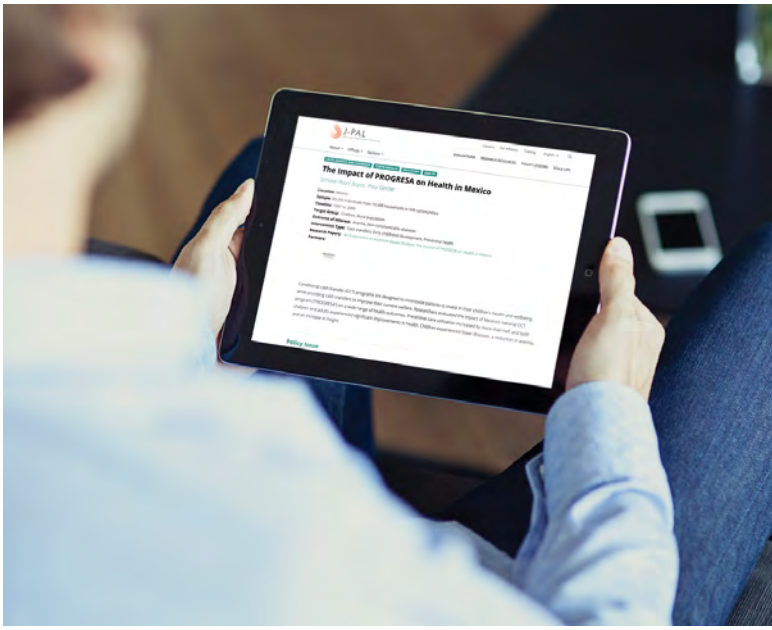


PHOTO: J-PAL

CASE STUDY: PUERTO RICO

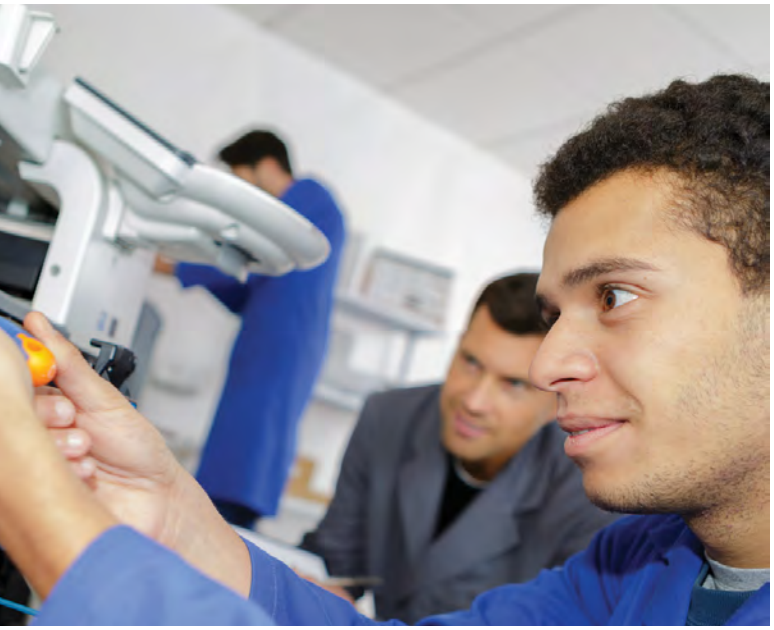


PHOTO: SHUTTERSTOCK.COM

How can a research project be sustained across an administration change?

Developing and carrying out a randomized evaluation often spans multiple years, so it is not uncommon for the researchers and policymakers collaborating on a randomized evaluation to have to adapt to an election and subsequent change in administration. In Puerto Rico, the process of designing a randomized evaluation to assess the impact of an earnings incentive and job-coaching program, called the Puerto Rico Self-Sufficiency Project, began under one governor's administration and now continues into another. By securing buy-in from staff at multiple levels, drawing on support from outside stakeholders, and providing opportunities for the new administration to provide input into the evaluation, the research team has been able to sustain the project through the administration change.

The Puerto Rico Self-Sufficiency Project aims to increase employment and earnings among current Temporary Assistance for Needy Families (TANF) participants by offering them a time-limited monthly financial incentive, conditional on formal sector employment. The project will also implement modified job-coaching and case management services that occur both before and after participants secure a job.

The evaluation will test the effect of the financial incentives alone, the modified job services alone, and the combined incentives and modified services on employment status, earnings, and social benefits payments to recipients. Researchers Gustavo J. Bobonis (Center for a New Economy, University of Toronto), Frederico Finan (University of California-Berkeley), Marco Gonzalez Navarro (University of Toronto), and Deepak Lamba Nieves (Center for a New Economy) are leading the study.

Securing buy-in from staff at all levels of the implementing agency early on eased the research projects' transition to the new administration after a new governor was elected in Puerto Rico in November 2016. Helping staff appreciate the value of the evaluation should be a priority regardless of upcoming elections, but also helps build sustainability across administrations.

In Puerto Rico, the turnover among political appointees in January 2017 did not imply the turnover of all staff contributing to the research project. Staff at multiple levels within the Administration for Socioeconomic Development of the Family (ADSEF), including non-appointees and service providers on the ground, had been engaged in developing the evaluation and during the transition. These continuing staff could attest that they understood, valued, and supported the evaluation—and wanted it to continue.

Partnering with a non-partisan, non-governmental organization, Espacios Abiertos, also played a pivotal role in easing the transition. Espacios Abiertos, an organization committed to increasing government transparency and civic engagement in Puerto Rico, is working with ADSEF to improve and implement the agency's job training program. As an organization based on the ground in Puerto Rico and not aligned with a particular party, Espacios Abiertos has been able to steward the evaluation through the governmental change.

Making the value of the evaluation clear to the incoming administration also helped sustain the research project. The randomized evaluation will provide actionable information to the government about whether the benefits of providing the earnings supplements and reforming the standard employment services outweigh their additional costs. The evaluation may provide additional potential value to the government by linking previously disparate

administrative datasets. The research team plans to link data on benefits payments collected by Puerto Rico's Department of the Family with data on employment and earnings collected by Puerto Rico's Department of Labor. This linkage could, for example, provide insights into labor force participation among TANF recipients in general.

Listening to the new administration and understanding its priorities has helped open the door to broader collaboration. Staff from J-PAL, the research team, and a co-chair of the State and Local Innovation Initiative visited Puerto Rico for an extensive series of meetings with the new administration. Much of these meetings focused not on introducing the specific evaluation of the Self Sufficiency Project, but rather on listening to the new administration. J-PAL staff and the researchers gathered information to track ongoing developments in the governmental shift—which departments were merging, who was coming in, what their priorities were, and how the project related to broader efforts. The meetings included not only ADSEF but also other officials and agencies, including the Secretary of State and the Secretary of Education.

By proactively looking for new connections after the administration change and being flexible about what the research project would yield, the research team and J-PAL have used the initial evaluation as an opportunity to help expand evidence-based policymaking efforts across a range of agencies. As one output from this engagement, J-PAL North America is hiring a Chief Evaluation Officer who will be embedded in the Department of Education. J-PAL Latin America & Caribbean will provide training opportunities to develop the government's capacity for research and evaluation.

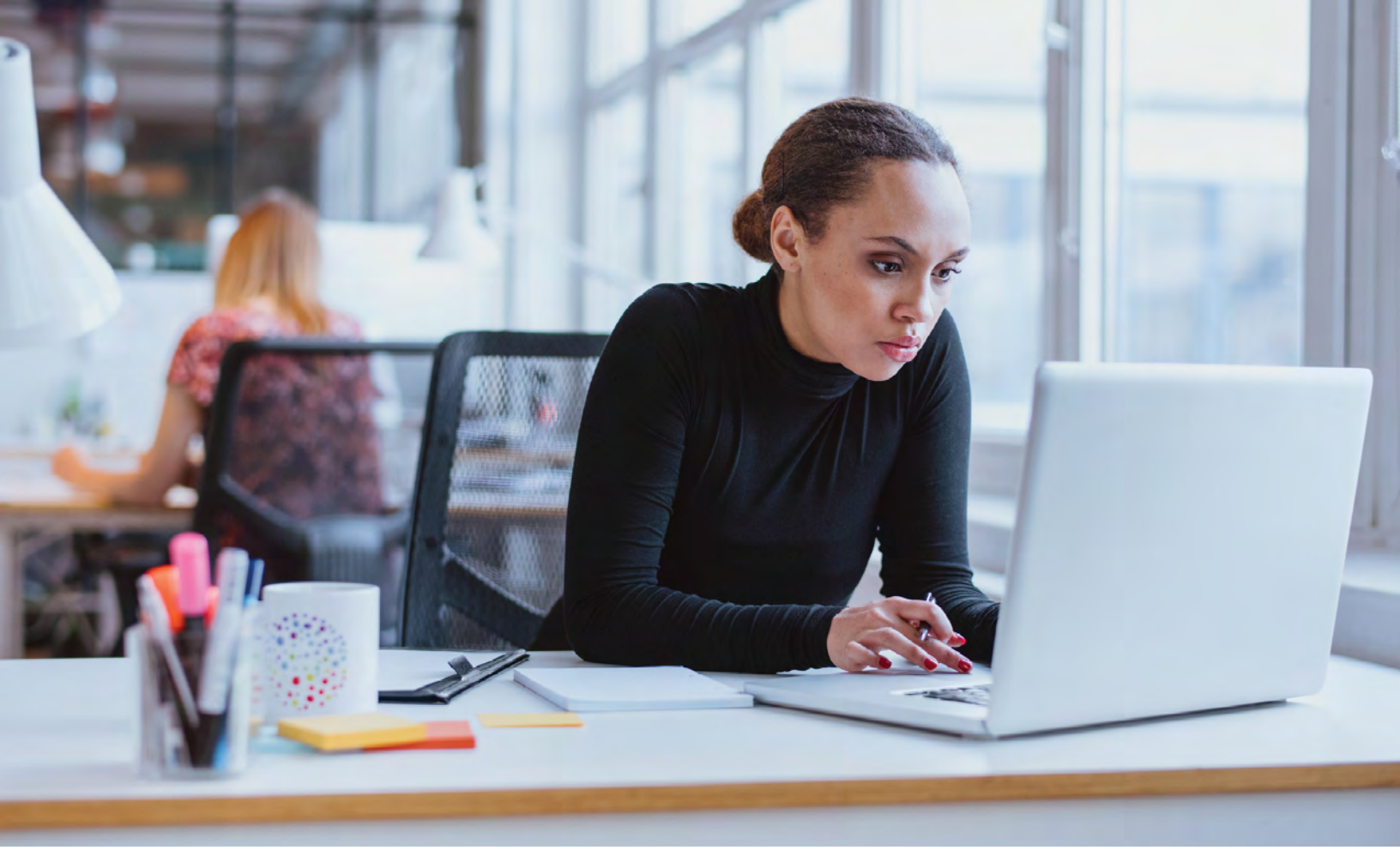


PHOTO: SHUTTERSTOCK.COM

CONCLUSION: MAKING FUTURE EVALUATIONS EASIER

The J-PAL State and Local Innovation Initiative aims to support state and local governments in launching randomized evaluations and to build their capacity to create and use rigorous evidence in the future. In our experience, state and local governments who have successfully partnered with academic researchers on a randomized evaluation are often well positioned to identify future opportunities for randomized evaluations and to assemble the internal and external resources and expertise needed to carry them out. Overall, the initiative aims to generate a shift in state and local policymakers' perceptions of the costs and benefits of randomized evaluations and to encourage further investments in evidence-based policymaking.

STRENGTHENING ADMINISTRATIVE DATA SYSTEMS

On page 17 of this guide, we discuss some of the potential benefits of using administrative data in randomized evaluations. Having access to administrative data that is high-quality, can be linked with other sources, and can be shared securely can enable evaluations that measure a range of potential outcomes at lower cost and with faster turnaround.

Despite these benefits, governments are often reluctant to share administrative data with researchers. This can be due to laws, regulations, or policies specifying which data can be shared. When government agencies are not sure which data they are permitted to share, they may err on the side of refusing all requests for data access. In other cases, governments lack the staff time and bandwidth to extract, prepare, and document which data are available.

Governments sometimes do not understand what data are available internally, which can also limit data sharing.

In some jurisdictions, these challenges cannot be easily overcome except by legislative action. Where legal restrictions continue to limit access to administrative data, researchers can collaborate with champions within government to highlight examples where research using administrative data has led to better program outcomes and reduced expenditures. At least some lawmakers have recognized these advantages and begun to advocate for improved data access.

However, governments can often make progress by devoting staff time and resources to strengthening and better understanding their own data systems. For instance, some governments have committed internal resources to cataloguing their data capabilities and promoting the existence of the resulting documentation, with the hope of inspiring more internal and external research projects. In the longer term, state and local governments can develop integrated data systems that gather data from across government (and sometimes private) agencies and link them using common identifiers.

Read more: Actionable Intelligence for Social Policy | bit.ly/2wuE3ly

CHANGING THE WAY RESEARCH IS FRAMED

In our experience, partnering with state and local governments to carry out randomized evaluations often involves building trust with stakeholders who have previously experienced evaluation as a thumbs-up or thumbs-down rating of a program's effectiveness. Frequently, this type of research is started at the behest of the federal government, funders, or other external parties. In contrast, when government stakeholders are able to play an active role in designing an evaluation with relevance for their own decision-making, it can change their perception of the value of rigorous evaluation.

Ensuring that the evidence generated by an evaluation is credible to decision-makers within government is key to shifting perceptions. One could likely find an anecdote, for example, supporting multiple contradictory views on the effectiveness of a program, making it very difficult to make decisions about the program. Having evidence from a randomized evaluation can lend clarity by shifting the discourse from questioning whether the evidence itself is sound, to questioning how best to interpret and apply the findings.

Additionally, governments can frame randomized evaluations not as a "one-off" but as part of a larger effort to improve their ability to address complex policy challenges. For example, the City of Philadelphia's ongoing evaluation of the WorkReady summer jobs program will ultimately look at the impact of the program on criminal justice, employment, and education outcomes. City leaders were also very interested in better understanding whether the program is reaching young people throughout the city, including in the most disadvantaged neighborhoods. The research team, with support from J-PAL North America staff, used linked program and administrative data to create detailed maps and analyses that provide insight into the young people served through the program and identify gaps in who is being served. In addition to providing useful information on how the program could improve targeting, this analysis helped build support for the randomized evaluation among key stakeholders within the City.

BUILDING MOMENTUM

Creative approaches developed by state and local governments and their research partners can overcome many of the challenges of launching a randomized evaluation. The lessons discussed throughout this guide highlight what we have learned from our partnerships with the governments selected to participate in the State and Local Innovation Initiative to date. Our hope is that these governments will serve as models for other state and local governments in the United States, demonstrating how to design high-quality and feasible randomized evaluations at the state and local level and encouraging others to consider randomized evaluations as a tool for addressing key challenges in their jurisdictions.

CASE STUDY: SOUTH CAROLINA

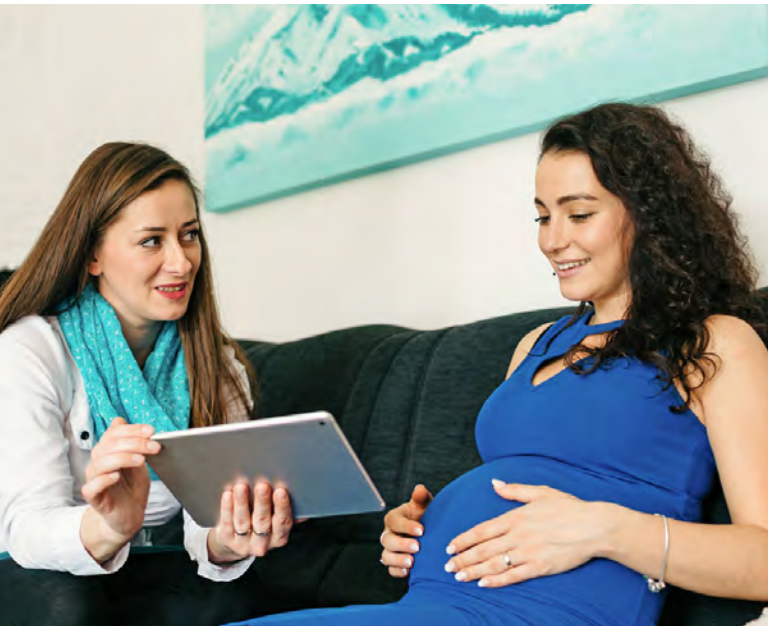


PHOTO: SHUTTERSTOCK.COM

How can one research project build momentum for evidence-based policymaking and make the next research project easier?

Designing and implementing a randomized evaluation requires close collaboration between government and research partners. Once that relationship has been formed, governments can draw upon the expertise of their research partners—and the expertise the government itself develops through an initial evaluation—to identify new opportunities and launch additional research projects with less time and effort. South Carolina's Department of Health and Human Services (HHS) exemplifies how a jurisdiction with a commitment to rigorous evaluation and using evidence to inform policy can leverage one study to catalyze a pipeline of evaluations.

Before participating in the State and Local Innovation Initiative, South Carolina had embarked on an expansion and evaluation of the Nurse-Family Partnership (NFP) home-visiting program, with the goal of improving maternal and child health. Through the program, specifically trained nurses visit low-income, first-time mothers regularly from early pregnancy through the child's second birthday, building trusted relationships in the process. The nurses advise and share expertise with the mothers to help them and their children achieve better health, well-being, and self-sufficiency. To expand the program statewide, South Carolina secured a Medicaid waiver to help cover the costs of home visits and established a pay-for-success contract. J-PAL North America is serving as the independent evaluator for the contract and is conducting a randomized evaluation to measure the impact of NFP on maternal and child health outcomes. The evaluation is currently being led by Katherine Baicker (University of Chicago Harris School School of Public Policy), Margaret McConnell (Harvard T.H. Chan School of Public Health), Mary Ann Bates (J-PAL North America Executive Director), Michelle Woodford (J-PAL North America Research Manager), and Annetta Zhou (Harvard University).

South Carolina's integrated data system, hosted by its Department of Revenue and Fiscal Affairs, facilitated the randomized evaluation. The integrated data system houses administrative data from multiple entities, including hospitals, emergency departments, and school districts. For the NFP evaluation, researchers will use linked data to understand the impact of NFP on a range of outcomes, beginning with pre-term births, birth spacing, and child injuries. Additionally, administrative data will allow researchers to track very long-term outcomes, such as the educational outcomes for the children of mothers who participate in the study.

The NFP evaluation brought many partners together and built experience within the state on how to leverage its integrated data system for research purposes to generate evidence and drive innovation. By the time South Carolina was selected to join the State and Local Innovation Initiative, the jurisdiction had already established a relationship with J-PAL North America, developed a concrete example of how its integrated data system could supply administrative data for a long-term evaluation, and created a vision for launching multiple research projects on key policy questions.

The starting point for South Carolina's engagement in the State and Local Innovation Initiative was tailored to fit the advanced stage of the partnership. Rather than focusing on introductions, J-PAL North America convened a kickoff meeting for agency leaders, including South Carolina's then-HHS Director, Christian Soura (now with the South Carolina Hospital Association), and researchers Craig Garthwaite (Northwestern University) and Matthew Notowidigdo (Northwestern University) to develop a list of potential research questions that could be explored over the year-long engagement. The preexisting relationship between

J-PAL North America and South Carolina facilitated an open, candid discussion of challenges the state faces—an evaluation pitching session Soura endearingly referred to as "J-PAL shark tank."

Because South Carolina had built a relationship not only with the NFP research team but also with J-PAL North America, the state gained an access point to multiple researchers through J-PAL's network. South Carolina could explore multiple projects outside of any single researcher's area of interest, and it could continue exploring projects even when individual researchers no longer had bandwidth to take on new projects. Ideas pitched included evaluations to measure the impact of Medicaid Managed Care Organizations (MCOs), medication-assisted treatment delivered via telemedicine, and opioid treatment interventions intended to reduce recidivism.

The most promising evaluation opportunity that emerged from these conversations aims to assess the impact of assigning Medicaid beneficiaries to different Managed Care Organizations. Historically, when individuals did not either actively choose an MCO or were not assigned to an MCO based on prior or family enrollment, they were then assigned to an MCO according to a round-robin algorithm. South Carolina worked with its Medicaid enrollment broker to shift from the quasi-random round-robin algorithm to a fully randomized process. Garthwaite and Notowidigdo will use historical Medicaid claims data and Medicaid claims data following implementation of the new random assignment protocol to estimate the relative impact of different MCOs on health care utilization.

Random assignment will allow them to distinguish the MCO's impact from any possible "cream skimming" effects—i.e., whether differences in health outcomes across MCOs is the result of different features of the MCOs, or the result of MCOs enrolling individuals who were more or less healthy to begin with. J-PAL North America's prior experience with South Carolina's integrated data system from the NFP evaluation made it easier to identify data sources for this project. Members of the NFP research team shared data dictionaries and insight on how to access health data with Garthwaite and Notowidigdo.

Beyond the MCO evaluation, South Carolina has continued working with J-PAL North America staff and researchers to identify future evaluation opportunities. As the year of technical support provided through the State and Local Innovation Initiative ended, J-PAL North America staff gave South Carolina a list of evaluation ideas of potential interest to both the state and J-PAL's affiliated researchers. Then-Director Soura weighed in on the list and identified ideas to continue exploring. Two ideas for evaluation have generated mutual interest between researchers and South Carolina, and these ongoing conversations hold potential to spur new research projects.

Overall, the time, effort, and social capital needed to pitch and scope each additional evaluation idea was reduced because a large investment in relationship building had already been made, and the state was already familiar with running a randomized evaluation. A clear takeaway from South Carolina is that research projects can build a pipeline such that as one project moves toward implementation, staff can leverage the existing relationship with researchers and external partners to begin scoping new projects.

APPENDIX: RESOURCES LISTED IN THIS GUIDE

Why Randomize? A one page primer on randomized evaluations:
bit.ly/2vlsXds

Common questions and concerns about randomized evaluations:
bit.ly/2im23Rm

Impact evaluation methods: what are they and what assumptions must hold for each to be valid?:
bit.ly/2rZ2O2p

Guide to asking the right research question:
bit.ly/2vUPfWV

Summary of the PROGRESA evaluation, health outcomes:
bit.ly/2imXtCB

Summary of the PROGRESA evaluation, school participation outcomes:
bit.ly/2wuqeTV

Summary of the SNAP take-up evaluation:
bit.ly/2wkpdgK

The risks of underpowered evaluations:
bit.ly/2vlEjDc

A summary of the Moving to Opportunity project:
bit.ly/2cpApxS

A list of resources on administrative data:
bit.ly/2wk8eeq

The Belmont Report on principles for the protection of human subjects of research (from The National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research)
bit.ly/2uvrnZO

The importance of piloting an RCT intervention (by Nancy Feeley, Sylvie Cossette, José Côté, Marjolaine Héon, Robyn Stremler, Geraldine Martorella, and Margaret Purden)
bit.ly/2xrX05z

The South Carolina Nurse Family Partnership pay for success pilot (from Social Finance in collaboration with Harvard T.H. Chan School of Public Health, J-PAL North America, Nurse-Family Partnership and the South Carolina Department of Health and Human Services)
bit.ly/2fLXCOA

The website for the group Actionable Intelligence for Social Policy, which helps governments with administrative data challenges:
bit.ly/2wuE3ly

BIBLIOGRAPHY

Attanasio, Orazio, Costas Meghir, and Ana Santiago. 2012. "Education Choices in Mexico: Using a Structural Model and a Randomized Experiment to evaluate Progresa." *The Review of Economic Studies* 79 (1): 37–66.

Chetty, Raj, Nathaniel Hendren, and Lawrence F. Katz. 2016. "The Effects of Exposure to Better Neighborhoods on Children: New Evidence from the Moving to Opportunity Experiment." *American Economic Review* 106 (4): 855-902.

Feeley, Nancy, Sylvie Cossette, José Côté, Marjolaine Héon, Robyn Stremler, Geraldine Martorella, and Margaret Purden. 2009. "The importance of piloting an RCT intervention." *Canadian Journal of Nursing Research* 41 (2): 84–99.

Finkelstein, Amy and Matthew Notowidigdo. 2016. "SNAP Take-Up Evaluation." J-PAL Evaluation Summary. https://www.povertyactionlab.org/evaluation/snap-take-evaluation.

Gelber, Alexander, Adam Isen and Judd B. Kessler. 2016. "The Effects of Youth Employment - Evidence from New York City Summer Youth Employment Program Lotteries." *Quarterly Journal of Economics* 131 (1): 423-460.

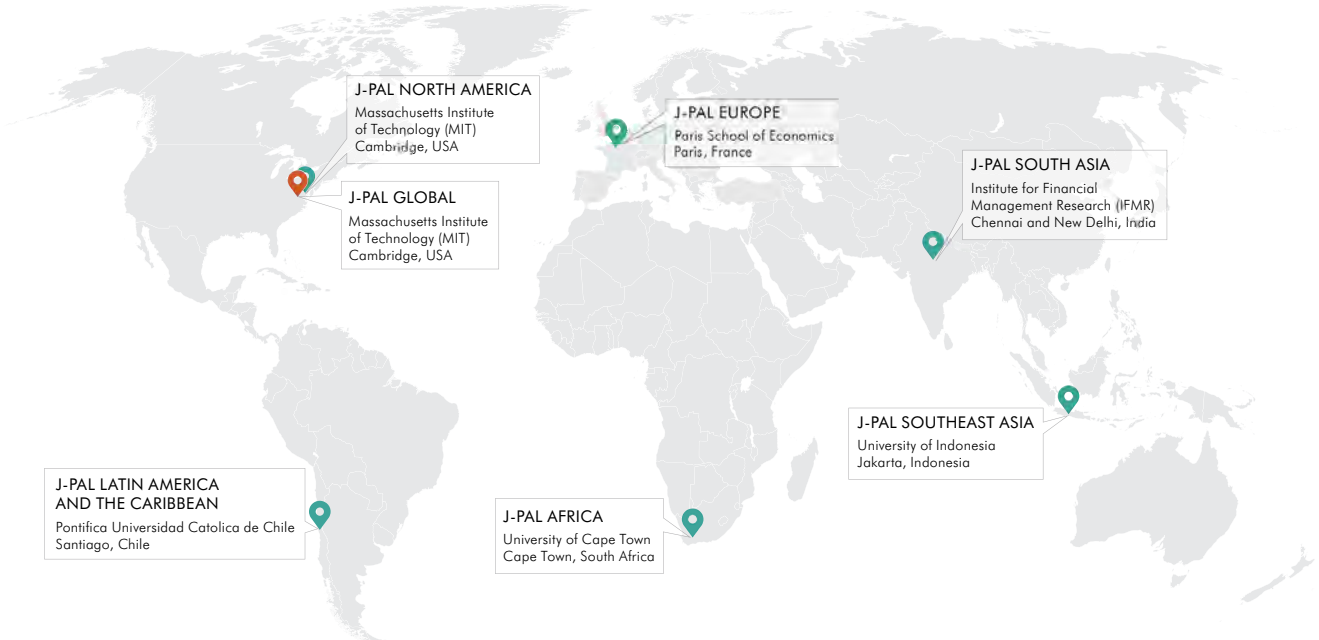
Gertler, Paul J., and Simone Boyce. 2001. "An Experiment in Incentive-Based Welfare: The Impact of PROGRESA on Health in Mexico." April 3. http://repec.org/res2003/Gertler.pdf.

Guryan, Jonathan, Sara Heller, Jens Ludwig, Sendhil Mullainathan, Harold Pollack, and Anuj Shah. 2017. "Thinking, Fast and Slow? Some Field Experiments to Reduce Crime and Dropout in Chicago." *The Quarterly Journal of Economics* 132 (1): 1-54.

Heller, Sara B. 2014. "Summer Jobs Reduce Violence Among Disadvantaged Youth." *Science* 346 (6214): 1219-1223.

Ludwig, Jens, Greg J. Duncan, Lisa A. Gennetian, Lawrence F. Katz, Ronald C. Kessler, Jeffrey R. Kling, and Lisa Sanbonmatsu. 2013. "Long-Term Neighborhood Effects on Low-Income Families: Evidence from Moving to Opportunity." *American Economic Review Papers & Proceedings* 103 (3): 226–31.

Ludwig, Jens, Lisa Sanbonmatsu, Lisa Gennetian, Emma Adam, Greg J. Duncan, Lawrence F. Katz, Ronald C. Kessler, Jeffrey R. Kling, Stacy Tessler Lindau, Robert C. Whitaker, and Thomas W. McDade. 2011. "Neighborhoods, Obesity, and Diabetes — A Randomized Social Experiment." *The New England Journal of Medicine* 365 (16): 1509–19.



J-PAL

ABDUL LATIF JAMEEL POVERTY ACTION LAB

NORTH AMERICA