Impact Evaluation in Practice

SECOND EDITION

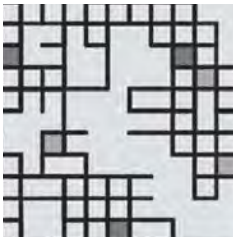
Paul J. Gertler, Sebastian Martinez,
Patrick Premand, Laura B. Rawlings,
and Christel M. J. Vermeersch

WORLD BANK GROUP

IDB
Inter-American Development Bank

Impact Evaluation in Practice

Second Edition

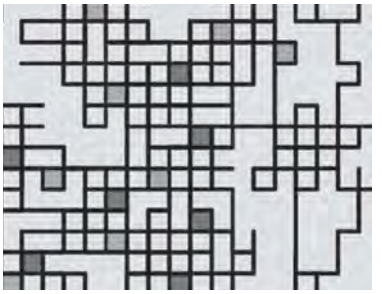


Please visit the *Impact Evaluation in Practice* book website at http://www.worldbank.org/ieinpractice. The website contains accompanying materials, including solutions to the book's HISP case study questions, as well as the corresponding data set and analysis code in the Stata software; a technical companion that provides a more formal treatment of data analysis; PowerPoint presentations related to the chapters; an online version of the book with hyperlinks to websites; and links to additional materials.

This book has been made possible thanks to the generous support of the Strategic Impact Evaluation Fund (SIEF). Launched in 2012 with support from the United Kingdom's Department for International Development, SIEF is a partnership program that promotes evidence-based policy making. The fund currently focuses on four areas critical to healthy human development: basic education, health systems and service delivery, early childhood development and nutrition, and water and sanitation. SIEF works around the world, primarily in low-income countries, bringing impact evaluation expertise and evidence to a range of programs and policy-making teams.

Impact Evaluation in Practice

Second Edition



Paul J. Gertler, Sebastian Martinez,
Patrick Premand, Laura B. Rawlings,
and Christel M. J. Vermeersch

WORLD BANK GROUP

IDB
Inter-American Development Bank

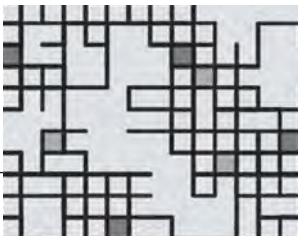
1818 H Street NW, Washington, DC 20433
Telephone: 202-473-1000; Internet: www.worldbank.org

1 2 3 4 19 18 17 16

ISBN (paper): 978-1-4648-0779-4
ISBN (electronic): 978-1-4648-0780-0
DOI: 10.1596/978-1-4648-0779-4

Illustration: C. Andres Gomez-Pena and Michaela Wieser
Cover Design: Critical Stages

Library of Congress Cataloging-in-Publication Data
Names: Gertler, Paul, 1955- author. | World Bank.
Title: Impact evaluation in practice / Paul J. Gertler, Sebastian Martinez, Patrick Premand, Laura B. Rawlings, Christel M. J. Vermeersch.
Description: Second Edition. | Washington, D.C.: World Bank, 2016. | Revised edition of Impact evaluation in practice, 2011.
Identifiers: LCCN 2016029061 (print) | LCCN 2016029464 (ebook) | ISBN 9781464807794 (pdf) | ISBN 9781464807800 | ISBN 9781464807800 ()
Subjects: LCSH: Economic development projects—Evaluation. | Evaluation research (Social action programs)
Classification: LCC HD75.9.G478 2016 (print) | LCC HD75.9 (ebook) | DDC 338.91—dc23
LC record available at https://lccn.loc.gov/2016029061

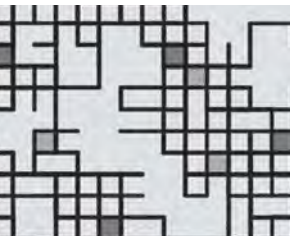


CONTENTS

Boxes

Figures

Tables



PREFACE

This book offers an accessible introduction to the topic of impact evaluation and its practice in development. It provides practical guidelines for designing and implementing impact evaluations, along with a nontechnical overview of impact evaluation methods.

This is the second edition of the *Impact Evaluation in Practice* handbook. First published in 2011, the handbook has been used widely by development and academic communities worldwide. The first edition is available in English, French, Portuguese, and Spanish.

The updated version covers the newest techniques for evaluating programs and includes state-of-the-art implementation advice, as well as an expanded set of examples and case studies that draw on recent development interventions. It also includes new material on research ethics and partnerships to conduct impact evaluation. Throughout the book, case studies illustrate applications of impact evaluations. The book links to complementary instructional material available online.

The approach to impact evaluation in this book is largely intuitive. We have tried to minimize technical notation. The methods are drawn directly from applied research in the social sciences and share many commonalities with research methods used in the natural sciences. In this sense, impact evaluation brings the empirical research tools widely used in economics and other social sciences together with the operational and political economy realities of policy implementation and development practice.

Our approach to impact evaluation is also pragmatic: we think that the most appropriate methods should be identified to fit the operational context, and not the other way around. This is best achieved at the outset of a program, through the design of prospective impact evaluations that are built into project implementation. We argue that gaining consensus among key stakeholders and identifying an evaluation design that fits the political

and operational context are as important as the method itself. We also believe that impact evaluations should be candid about their limitations and caveats. Finally, we strongly encourage policy makers and program managers to consider impact evaluations as part of a well-developed theory of change that clearly sets out the causal pathways by which a program works to produce outputs and influence final outcomes, and we encourage them to combine impact evaluations with monitoring and complementary evaluation approaches to gain a full picture of results.

Our experiences and lessons on how to do impact evaluation in practice are drawn from teaching and working with hundreds of capable government, academic, and development partners. The book draws, collectively, from dozens of years of experience working with impact evaluations in almost every corner of the globe and is dedicated to future generations of practitioners and policy makers.

We hope the book will be a valuable resource for the international development community, universities, and policy makers looking to build better evidence around what works in development. More and better impact evaluations will help strengthen the evidence base for development policies and programs around the world. Our hope is that if governments and development practitioners can make policy decisions based on evidence—including evidence generated through impact evaluation—development resources will be spent more effectively to reduce poverty and improve people's lives.

Road Map to Contents of the Book

Part 1–Introduction to Impact Evaluation (chapters 1 and 2) discusses why an impact evaluation might be undertaken and when it is worthwhile to do so. We review the various objectives that an impact evaluation can achieve and highlight the fundamental policy questions that an evaluation can tackle. We insist on the necessity of carefully tracing a theory of change that explains the channels through which programs can influence final outcomes. We urge careful consideration of outcome indicators and anticipated effect sizes.

Part 2–How to Evaluate (chapters 3 through 10) reviews various methodologies that produce comparison groups that can be used to estimate program impacts. We begin by introducing the *counterfactual* as the crux of any impact evaluation, explaining the properties that the estimate of the counterfactual must have, and providing examples of invalid estimates of the counterfactual. We then present a menu of impact evaluation options that can produce valid estimates of the counterfactual. In particular,

we discuss the basic intuition behind five impact evaluation methodologies: *randomized assignment, instrumental variables, regression discontinuity design, difference-in-differences,* and *matching*. We discuss why and how each method can produce a valid estimate of the counterfactual, in which policy context each can be implemented, and the main limitations of each method.

Throughout this part of the book, a case study—the Health Insurance Subsidy Program (HISP)—is used to illustrate how the methods can be applied. In addition, we present specific examples of impact evaluations that have used each method. Part 2 concludes with a discussion of how to combine methods and address problems that can arise during implementation, recognizing that impact evaluation designs are often not implemented exactly as originally planned. In this context, we review common challenges encountered during implementation, including imperfect compliance or spillovers, and discuss how to address these issues. Chapter 10 concludes with guidance on evaluations of multifaceted programs, notably those with different treatment levels and crossover designs.

Part 3–How to Implement an Impact Evaluation (chapters 11 through 14) focuses on how to implement an impact evaluation, beginning in chapter 11 with how to use the rules of program operation—namely, a program's available resources, criteria for selecting beneficiaries, and timing for implementation—as the basis for selecting an impact evaluation method. A simple framework is set out to determine which of the impact evaluation methodologies presented in part 2 is most suitable for a given program, depending on its operational rules. Chapter 12 discusses the relationship between the research team and policy team and their respective roles in jointly forming an evaluation team. We review the distinction between independence and unbiasedness, and highlight areas that may prove to be sensitive in carrying out an impact evaluation. We provide guidance on how to manage expectations, highlight some of the common risks involved in conducting impact evaluations, and offer suggestions on how to manage those risks. The chapter concludes with an overview of how to manage impact evaluation activities, including setting up the evaluation team, timing the evaluation, budgeting, fundraising, and collecting data. Chapter 13 provides an overview of the ethics and science of impact evaluation, including the importance of not denying benefits to eligible beneficiaries for the sake of the evaluation; outlines the role of institutional review boards that approve and monitor research involving human subjects; and discusses the importance of registering evaluations following the practice of open science, whereby data are made publicly available for further research and for replicating results. Chapter 14 provides insights into how to use impact

evaluations to inform policy, including tips on how to make the results relevant; a discussion of the kinds of products that impact evaluations can and should deliver; and guidance on how to produce and disseminate findings to maximize policy impact.

Part 4–How to Get Data for an Impact Evaluation (chapters 15 through 17) discusses how to collect data for an impact evaluation, including choosing the sample and determining the appropriate size of the evaluation sample (chapter 15), as well as finding adequate sources of data (chapter 16). Chapter 17 concludes and provides some checklists.

Complementary Online Material

Accompanying materials are located on the Impact Evaluation in Practice website (http://www.worldbank.org/ieinpractice), including solutions to the book's HISP case study questions, the corresponding data set and analysis code in the Stata software, as well as a technical companion that provides a more formal treatment of data analysis. Materials also include PowerPoint presentations related to the chapters, an online version of the book with hyperlinks to websites, and links to additional materials.

The Impact Evaluation in Practice website also links to related material from the World Bank Strategic Impact Evaluation Fund (SIEF), Development Impact Evaluation (DIME), and Impact Evaluation Toolkit websites, as well as the Inter-American Development Bank Impact Evaluation Portal and the applied impact evaluation methods course at the University of California, Berkeley.

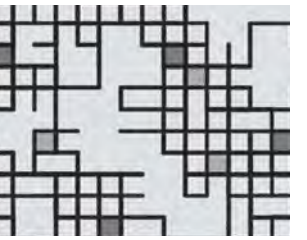
Development of *Impact Evaluation in Practice*

The first edition of the *Impact Evaluation in Practice* book built on a core set of teaching materials developed for the "Turning Promises to Evidence" workshops organized by the Office of the Chief Economist for Human Development, in partnership with regional units and the Development Economics Research Group at the World Bank. At the time of writing the first edition, the workshop had been delivered more than 20 times in all regions of the world.

The workshops and both the first and second editions of this handbook have been made possible thanks to generous grants from the Spanish government, the United Kingdom's Department for International Development (DFID), and the Children's Investment Fund Foundation (CIFF UK),

through contributions to the Strategic Impact Evaluation Fund (SIEF). The second edition has also benefited from support from the Office of Strategic Planning and Development Effectiveness at the Inter-American Development Bank (IDB).

This second edition has been updated to cover the most up-to-date techniques and state-of-the-art implementation advice following developments made in the field in recent years. We have also expanded the set of examples and case studies to reflect wide-ranging applications of impact evaluation in development operations and underline its linkages to policy. Lastly, we have included applications of impact evaluation techniques with Stata, using the HISP case study data set, as part of the complementary online material.



ACKNOWLEDGMENTS

The teaching materials on which the book is based have been through numerous incarnations and have been taught by a number of talented faculty, all of whom have left their mark on the methods and approach to impact evaluation espoused in the book. We would like to thank and acknowledge the contributions and substantive input of a number of faculty who have co-taught the workshops on which the first edition was built, including Paloma Acevedo Alameda, Felipe Barrera, Sergio Bautista-Arredondo, Stefano Bertozzi, Barbara Bruns, Pedro Carneiro, Jishnu Das, Damien de Walque, David Evans, Claudio Ferraz, Deon Filmer, Jed Friedman, Emanuela Galasso, Sebastian Galiani, Arianna Legovini, Phillippe Leite, Gonzalo Hernández Licona, Mattias Lundberg, Karen Macours, Juan Muñoz, Plamen Nikolov, Berk Özler, Nancy Qian, Gloria M. Rubio, Norbert Schady, Julieta Trias, and Sigrid Vivo Guzman. We are grateful for comments from our peer reviewers for the first edition of the book (Barbara Bruns, Arianna Legovini, Dan Levy, and Emmanuel Skoufias) and the second edition (David Evans, Francisco Gallego, Dan Levy, and Damien de Walque), as well as from Gillette Hall. We also gratefully acknowledge the efforts of a talented workshop organizing team, including Holly Balgrave, Theresa Adobea Bampoe, Febe Mackey, Silvia Paruzzolo, Tatyana Ringland, Adam Ross, and Jennifer Sturdy.

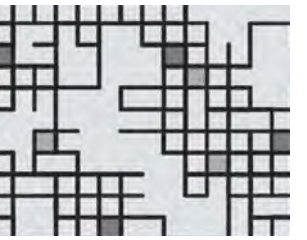
We thank all the individuals who participated in drafting transcripts of the July 2009 workshop in Beijing, China, on which parts of this book are based, particularly Paloma Acevedo Alameda, Carlos Asenjo Ruiz, Sebastian Bauhoff, Bradley Chen, Changcheng Song, Jane Zhang, and Shufang Zhang. We thank Garret Christensen and the Berkeley Initiative for Transparency in the Social Sciences, as well as Jennifer Sturdy and Elisa Rothenbühler, for inputs to chapter 13. We are also grateful to Marina Tolchinsky and Kristine Cronin for excellent research assistance; Cameron Breslin and Restituto Cardenas for scheduling support; Marco Guzman and Martin

Ruegenberg for designing the illustrations; and Nancy Morrison, Cindy A. Fisher, Fiona Mackintosh, and Stuart K. Tucker for editorial support during the production of the first and second editions of the book.

We gratefully acknowledge the continued support and enthusiasm for this project from our managers at the World Bank and Inter-American Development Bank, and especially from the SIEF team, including Daphna Berman, Holly Blagrave, Restituto Cardenas, Joost de Laat, Ariel Fiszbein, Alaka Holla, Aliza Marcus, Diana-Iuliana Pirjol, Rachel Rosenfeld, and Julieta Trias. We are very grateful for the support received from SIEF management, including Luis Benveniste, Joost de Laat, and Julieta Trias. We are also grateful to Andrés Gómez-Peña and Michaela Wieser from the Inter-American Development Bank and Mary Fisk, Patricia Katayama, and Mayya Revzina from the World Bank for their assistance with communications and the publication process.

Finally, we would like to thank the participants in numerous workshops, notably those held in Abidjan, Accra, Addis Ababa, Amman, Ankara, Beijing, Berkeley, Buenos Aires, Cairo, Cape Town, Cuernavaca, Dakar, Dhaka, Fortaleza, Kathmandu, Kigali, Lima, Madrid, Managua, Manila, Mexico City, New Delhi, Paipa, Panama City, Pretoria, Rio de Janeiro, San Salvador, Santiago, Sarajevo, Seoul, Sofia, Tunis, and Washington, DC.

Through their interest, sharp questions, and enthusiasm, we were able to learn step by step what policy makers are looking for in impact evaluations. We hope this book reflects their ideas.



ABOUT THE AUTHORS

Paul J. Gertler is the Li Ka Shing Professor of Economics at the University of California at Berkeley, where he holds appointments in the Haas School of Business and the School of Public Health. He is also the Scientific Director of the University of California Center for Effective Global Action. He was Chief Economist of the Human Development Network of the World Bank from 2004 to 2007 and the Founding Chair of the Board of Directors of the International Initiative for Impact Evaluation (3ie) from 2009 to 2012. At the World Bank, he led an effort to institutionalize and scale up impact evaluation for learning what works in human development. He has been a Principal Investigator on a large number of at-scale multisite impact evaluations including Mexico's CCT program, PROGRESA/OPORTUNIDADES, and Rwanda's Health Care Pay-for-Performance scheme. He holds a PhD in economics from the University of Wisconsin and has held academic appointments at Harvard, RAND, and the State University of New York at Stony Brook.

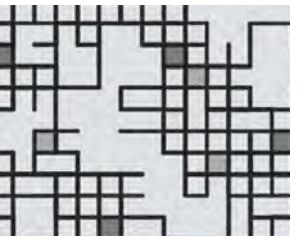
Sebastian Martinez is a Principal Economist in the Office of Strategic Planning and Development Effectiveness at the Inter-American Development Bank (IDB). His work focuses on strengthening the evidence base and development effectiveness of the social and infrastructure sectors, including health, social protection, labor markets, water and sanitation, and housing and urban development. He heads a team of economists that conducts research on the impacts of development programs and policies, supports the implementation of impact evaluations for operations, and conducts capacity development for clients and staff. Prior to joining the IDB, he spent six years at the World Bank, leading evaluations of social programs in Latin America and Sub-Saharan Africa. He holds a PhD in economics from the University of California at Berkeley, with a specialization in development and applied microeconomics.

Patrick Premand is a Senior Economist in the Social Protection and Labor Global Practice at the World Bank. He conducts analytical and operational work on social protection and safety nets; labor markets, youth employment and entrepreneurship; as well as early childhood development. His research focuses on building evidence on the effectiveness of development policies through impact evaluations of large-scale social and human development programs. He previously held various other positions at the World Bank, including in the Human Development Economics Unit of the Africa region, the Office of the Chief Economist for Human Development, and the Poverty Unit of the Latin America and the Caribbean region. He holds a DPhil in economics from Oxford University.

Laura B. Rawlings is a Lead Social Protection Specialist at the World Bank, with over 20 years of experience in the design, implementation, and evaluation of human development programs. She manages both operations and research, with a focus on developing innovative approaches for effective, scalable social protection systems in low-resource settings. She was the team leader responsible for developing the World Bank's Social Protection and Labor Strategy 2012–22 and was previously the manager of the Strategic Impact Evaluation Fund (SIEF). She also worked as the Sector Leader for Human Development in Central America, where she was responsible for managing the World Bank's health, education, and social protection portfolios. She began her career at the World Bank in the Development Research Group, where she worked on the impact evaluation of social programs. She has worked in Latin America and the Caribbean as well as Sub-Saharan Africa, leading numerous project and research initiatives in the areas of conditional cash transfers, public works, social funds, early childhood development, and social protection systems. Prior to joining the World Bank, she worked for the Overseas Development Council, where she ran an education program on development issues for staff in the United States Congress. She has published numerous books and articles in the fields of evaluation and human development and is an adjunct professor in the Global Human Development program at Georgetown University, Washington DC.

Christel M. J. Vermeersch is a Senior Economist in the Health, Nutrition and Population Global Practice at the World Bank. She works on issues related to health sector financing, results-based financing, monitoring and evaluation, and impact evaluation. She previously worked in the education, early childhood development, and skills areas. She has coauthored impact evaluation studies for results-based financing programs in Argentina and

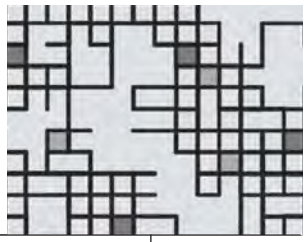
Rwanda, a long-term follow-up of an early childhood stimulation study in Jamaica, as well as the World Bank's impact evaluation toolkit for health. Prior to joining the World Bank, she was a Prize Postdoctoral Research Fellow at Oxford University. She holds a PhD in economics from Harvard University.



ABBREVIATIONS

3IE	International Initiative for Impact Evaluation
ATE	average treatment effect
CCT	conditional cash transfer
CITI	Collaborative Institutional Training Initiative
DD	difference-in-differences, or double differences
DIME	Development Impact Evaluation (World Bank)
HISP	Health Insurance Subsidy Program
ID	identification number
IDB	Inter-American Development Bank
IHSN	International Household Survey Network
IRB	institutional review board
ITT	intention-to-treat
IV	instrumental variables
J-PAL	Abdul Latif Jameel Poverty Action Lab
LATE	local average treatment effect
MDE	minimum detectable effect
NGO	nongovernmental organization
NIH	National Institutes of Health (United States)
ODI	Overseas Development Institute
OSF	Open Science Framework
RCT	randomized controlled trial
RDD	regression discontinuity design
RIDIE	Registry for International Development Impact Evaluations

SIEF	Strategic Impact Evaluation Fund (World Bank)
SMART	specific, measurable, attributable, realistic, and targeted
SUTVA	stable unit treatment value assumption
TOT	treatment-on-the-treated
UN	United Nations
USAID	United States Agency for International Development
WHO	World Health Organization



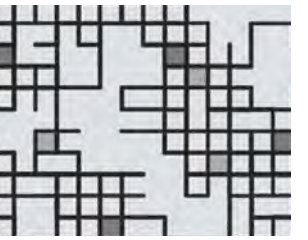
Part 1

INTRODUCTION TO IMPACT EVALUATION

The first part of the book presents an overview of impact evaluation. Chapter 1 discusses why impact evaluation is important and how it fits within the context of ethical, evidence-based policy making. We contrast impact evaluation with monitoring, introduce the defining features of impact evaluation, and discuss complementary approaches, including cost-benefit analysis and cost-effectiveness analysis. We introduce a core focus of the book: namely, how a program's available resources, eligibility criteria for selecting beneficiaries, and timing for implementation serve to structure options in the selection of impact evaluation methods. Finally, we introduce different modalities of impact evaluation—such as prospective and retrospective evaluation, and efficacy versus effectiveness trials—and conclude with a discussion on when to use impact evaluations.

Chapter 2 discusses how to formulate evaluation questions and hypotheses that are useful for policy. These questions and hypotheses determine

the focus of the evaluation. We also introduce the fundamental concept of a theory of change and the related use of results chains and performance indicators. Chapter 2 provides the first introduction to the fictional case study, the Health Insurance Subsidy Program (HISP), that is used throughout the book and in the accompanying material found on the Impact Evaluation in Practice website (www.worldbank.org/ieinpractice).



CHAPTER 1

Why Evaluate?

Evidence-Based Policy Making

Development programs and policies are typically designed to change outcomes such as raising incomes, improving learning, or reducing illness. Whether or not these changes are actually achieved is a crucial public policy question, but one that is not often examined. More commonly, program managers and policy makers focus on measuring and reporting the inputs and immediate outputs of a program—how much money is spent, how many textbooks are distributed, how many people participate in an employment program—rather than on assessing whether programs have achieved their intended goals of improving outcomes.

Impact evaluations are part of a broader agenda of *evidence-based policy making*. This growing global trend is marked by a shift in focus from inputs to outcomes and results, and is reshaping public policy. Not only is the focus on results being used to set and track national and international targets, but results are increasingly being used by, and required of, program managers to enhance accountability, determine budget allocations, and guide program design and policy decisions.

Monitoring and evaluation are at the heart of evidence-based policy making. They provide a core set of tools that stakeholders can use to verify and improve the quality, efficiency, and effectiveness of policies and programs at various stages of implementation—or, in other words, to focus on results. At the program management level, there is a need to

understand which program design options are most cost-effective, or make the case to decision makers that programs are achieving their intended results in order to obtain budget allocations to continue or expand them. At the country level, ministries compete with one another to obtain funding from the ministry of finance. And finally, governments are accountable to citizens to inform them of the performance of public programs. Evidence can constitute a strong foundation for transparency and accountability.

The robust evidence generated by impact evaluations is increasingly serving as a foundation for greater accountability, innovation, and learning. In a context in which policy makers and civil society are demanding results and accountability from public programs, impact evaluation can provide robust and credible evidence on performance and, crucially, on whether a particular program has achieved or is achieving its desired outcomes. Impact evaluations are also increasingly being used to test innovations in program design or service delivery. At the global level, impact evaluations are central to building knowledge about the effectiveness of development programs by illuminating what does and does not work to reduce poverty and improve welfare.

Simply put, an impact evaluation assesses the changes in the well-being of individuals that can be attributed to a particular project, program, or policy. This focus on attribution is the hallmark of impact evaluations. Correspondingly, the central challenge in carrying out effective impact evaluations is to identify the causal relationship between the program or policy and the outcomes of interest.

Impact evaluations generally estimate average impacts of a program, program modalities, or a design innovation. For example, did a water and sanitation program increase access to safe water and improve health outcomes? Did a new curriculum raise test scores among students? Was the innovation of including noncognitive skills as part of a youth training program successful in fostering entrepreneurship and raising incomes? In each of these cases, the impact evaluation provides information on whether the program caused the desired changes in outcomes, as contrasted with specific case studies or anecdotes, which can give only partial information and may not be representative of overall program impacts. In this sense, well-designed and well-implemented impact evaluations are able to provide convincing and comprehensive evidence that can be used to inform policy decisions, shape public opinion, and improve program operations.

Classic impact evaluations address the effectiveness of a program against the absence of the program. Box 1.1 covers the well-known impact evaluation of Mexico's conditional cash transfer (CCT) program,

Box 1.1: How a Successful Evaluation Can Promote the Political Sustainability of a Development Program: Mexico's Conditional Cash Transfer Program

In the 1990s, the government of Mexico launched an innovative conditional cash transfer (CCT) program first called Progresa (the name changed, together with a few elements of the program, to Oportunidades, and then to Prospera). Its objectives were to provide poor households with short-term income support and to create incentives for investments in children's human capital, primarily by providing cash transfers to mothers in poor households conditional on their children regularly attending school and visiting a health center.

From the beginning, the government considered it essential to monitor and evaluate the program. The program's officials contracted a group of researchers to design an impact evaluation and build it into the program's expansion at the same time that it was rolled out successively to the participating communities.

The 2000 presidential election led to a change of the party in power. In 2001, Progresa's external evaluators presented their findings to the newly elected administration. The results of the program were impressive: they showed that the program was well targeted to the poor and had engendered promising changes in households' human capital. Schultz (2004) found that the program significantly improved school enrollment, by an average of 0.7 additional years of schooling. Gertler (2004) found that the incidence of illness in children decreased by 23 percent, while the number of sick or disability days fell by 19 percent among adults. Among the nutritional outcomes, Behrman and Hoddinott (2001) found that the program reduced the probability of stunting by about 1 centimeter per year for children in the critical age range of 12–36 months.

These evaluation results supported a political dialogue based on evidence and contributed to the new administration's decision to continue the program. The government expanded the program's reach, introducing upper-middle school scholarships and enhanced health programs for adolescents. At the same time, the results were used to modify other social assistance programs, such as the large and less well-targeted tortilla subsidy, which was scaled back.

The successful evaluation of Progresa also contributed to the rapid adoption of CCTs around the world, as well as Mexico's adoption of legislation requiring all social projects to be evaluated.

Sources: Behrman and Hoddinott 2001; Fiszbein and Schady 2009; Gertler 2004; Levy and Rodríguez 2005; Schultz 2004; Skoufias and McClafferty 2001.

illustrating how the evaluation contributed to policy discussions concerning the expansion of the program.[1]

Box 1.2 illustrates how impact evaluation influenced education policy in Mozambique by showing that community-based preschools can be an affordable and effective way to address early education and prompt children to enroll in primary school at the right age.

In addition to addressing the basic question of whether a program is effective or not, impact evaluations can also be used to explicitly test alternative program modalities or design innovations. As policy makers become increasingly focused on better understanding how to improve implementation and gain value for money, approaches testing design alternatives are rapidly gaining ground. For example, an evaluation might compare the performance of a training program to that of a promotional campaign to

Box 1.2: The Policy Impact of an Innovative Preschool Model: Preschool and Early Childhood Development in Mozambique

While preschool is recognized as a good investment and effective approach to preparing children for school and later life, developing countries have struggled with the question of how to introduce a scalable and cost-effective preschool model. In Mozambique, only about 4 percent of children attend preschool. Upon reaching primary school, some children from rural communities show signs of developmental delays and are often not prepared for the demands of the education system. Moreover, despite the primary school enrollment rate of nearly 95 percent, one-third of children are not enrolled by the appropriate age.

In 2006, Save the Children piloted a community-based preschool program in rural communities of Mozambique aiming to improve children's cognitive, social, emotional, and physical development. In what is believed to be the first randomized evaluation of a preschool program in rural Africa, a research team conducted an impact evaluation of the program in 2008. Based on the evaluation's positive results, the government of Mozambique adopted and decided to expand Save the Children's community-based preschool model to 600 communities.

The evaluation found that children who attended preschool were 24 percent more likely to enroll in primary school and 10 percent more likely to start at the appropriate age than children in the comparison group. In primary school, children who had attended preschool spent almost 50 percent more time on homework and other school-related activities than those who did not. The evaluation also showed positive gains in school readiness; children who attended preschool performed better on tests of cognitive, socioemotional, and fine motor development in comparison to the comparison group.

Other household members also benefitted from children's enrollment in preschool by having more time to engage in productive activities. Older siblings were 6 percent more likely to attend school and caregivers were 26 percent more likely to have worked in the previous 30 days when a young child in the household attended preschool.

This evaluation showed that even in a low-income setting, preschools can be an effective way to foster cognitive development, prepare children for primary school, and increase the likelihood that children will begin primary school at the appropriate age.

Source: Martinez, Nadeau, and Pereira 2012.

see which one is more effective in raising financial literacy. An impact evaluation can test which combination of nutrition and child stimulation approaches has the largest impact on child development. Or the evaluation might test a design innovation to improve an existing program, such as using text messages to prompt compliance with taking prescribed medications.

What Is Impact Evaluation?

Impact evaluation is one of many approaches that support evidence-based policy, including monitoring and other types of evaluation.

Monitoring is a continuous process that tracks what is happening within a program and uses the data collected to inform program implementation and day-to-day management and decisions. Using mostly administrative data, the process of monitoring tracks financial disbursement and program performance against expected results, and analyzes trends over time.[2] Monitoring is necessary in all programs and is a critical source of information about program performance, including implementation and costs. Usually, monitoring tracks inputs, activities, and outputs, although occasionally it can include outcomes, such as progress toward achieving national development goals.

Evaluations are periodic, objective assessments of a planned, ongoing, or completed project, program, or policy. Evaluations are used selectively to answer specific questions related to design, implementation, and results. In contrast to continuous monitoring, they are carried out at discrete points in time and often seek an outside perspective from technical experts. Their design, method, and cost vary substantially depending on the type of question the evaluation is trying to answer. Broadly speaking, evaluations can address three types of questions (Imas and Rist 2009):[3]

Key Concept

Evaluations are periodic, objective assessments of a planned, ongoing, or completed project, program, or policy. Evaluations are used to answer specific questions, often related to design, implementation, or results.

- *Descriptive questions* ask about what is taking place. They are concerned with processes, conditions, organizational relationships, and stakeholder views.
- *Normative questions* compare what is taking place to what should be taking place. They assess activities and whether or not targets are accomplished. Normative questions can apply to inputs, activities, and outputs.
- *Cause-and-effect questions* focus on attribution. They ask about what difference the intervention makes to outcomes.

There are many types of evaluations and evaluation methods, drawing on both quantitative and qualitative data. *Qualitative data* are expressed not in numbers, but rather by means of language or sometimes images. *Quantitative data* are numerical measurements and are commonly associated with scales or metrics. Both quantitative and qualitative data can be used to answer the types of questions posed above. In practice, many evaluations rely on both types of data. There are multiple data sources that can be used for evaluations, drawing on primary data collected for the purpose of the evaluation or available secondary data (see chapter 16 on data sources). This book focuses on impact evaluations using quantitative data, but underscores the value of monitoring, of complementary evaluation methods, and of using both quantitative and qualitative data.

Key Concept

Impact evaluations seek to answer one particular type of question: What is the impact (or causal effect) of a program on an outcome of interest?

Impact evaluations are a particular type of evaluation that seeks to answer a specific cause-and-effect question: What is the impact (or causal effect) of a program on an outcome of interest? This basic question incorporates an important causal dimension. The focus is only on the *impact*: that is, the changes *directly attributable* to a program, program modality, or design innovation.

The basic evaluation question—what is the impact or causal effect of a program on an outcome of interest?—can be applied to many contexts. For instance, what is the causal effect of scholarships on school attendance and academic achievement? What is the impact of contracting out primary care to private providers on access to health care? If dirt floors are replaced with cement floors, what will be the impact on children's health? Do improved roads increase access to labor markets and raise households' income, and if so, by how much? Does class size influence student achievement, and if it does, by how much? As these examples show, the basic evaluation question can be extended to examine the impact of a *program modality or design innovation,* not just a program.

The focus on causality and attribution is the hallmark of impact evaluations. All impact evaluation methods address some form of *cause-and-effect* question. The approach to addressing causality determines the methodologies that can be used. To be able to estimate the causal effect or impact of a program on outcomes, any impact evaluation method chosen must estimate the so-called *counterfactual*: that is, what the outcome would have been for program participants if they had not participated in the program. In practice, impact evaluation requires that the evaluation team find a comparison group to estimate what would have happened to the program participants without the program, then make comparisons with the treatment group that has received the program. Part 2 of the

book describes the main methods that can be used to find adequate comparison groups.

One of the main messages of this book is that the choice of an impact evaluation method depends on the operational characteristics of the program being evaluated. When the rules of program operation are equitable and transparent and provide accountability, a good impact evaluation design can almost always be found—provided that the impact evaluation is planned early in the process of designing or implementing a program. Having clear and well-defined rules of program operations not only has intrinsic value for sound public policy and program management, it is also essential for constructing good comparison groups—the foundation of rigorous impact evaluations. Specifically, the choice of an impact evaluation method is determined by the operational characteristics of the program, notably its available resources, eligibility criteria for selecting beneficiaries, and timing for program implementation. As we will discuss in parts 2 and 3 of the book, you can ask three questions about the operational context of a given program: Does your program have resources to serve all eligible beneficiaries? Is your program targeted or universal? Will your program be rolled out to all beneficiaries at once or in sequence? The answer to these three questions will determine which of the methods presented in part 2—randomized assignment, instrumental variables, regression discontinuity, difference-in-differences, or matching—are the most suitable to your operational context.

Key Concept

The choice of an impact evaluation method depends on the operational characteristics of the program being evaluated, notably its available resources, eligibility criteria for selecting beneficiaries, and timing for program implementation.

Prospective versus Retrospective Impact Evaluation

Impact evaluations can be divided into two categories: prospective and retrospective. *Prospective evaluations* are developed at the same time as the program is being designed and are built into program implementation. Baseline data are collected before the program is implemented for both the group receiving the intervention (known as the *treatment group*) and the group used for comparison that is not receiving the intervention (known as the *comparison group*). *Retrospective evaluations* assess program impact after the program has been implemented, looking for treatment and comparison groups ex post.

Key Concept

Prospective evaluations are designed and put in place before a program is implemented.

Prospective impact evaluations are more likely to produce strong and credible evaluation results, for three reasons. First, baseline data can be collected to establish measures of outcomes of interest before the program has started. Baseline data are important for measuring

pre-intervention outcomes. Baseline data on the treatment and comparison groups should be analyzed to ensure that the groups are similar. Baselines can also be used to assess targeting effectiveness: that is, whether or not the program is reaching its intended beneficiaries.

Second, defining measures of a program's success in the program's planning stage focuses both the program and the evaluation on intended results. As we shall see, impact evaluations take root in a program's theory of change or results chain. The design of an impact evaluation helps clarify program objectives—particularly because it requires establishing well-defined measures of a program's success. Policy makers should set clear goals for the program to meet, and clear questions for the evaluation to answer, to ensure that the results will be highly relevant to policy. Indeed, the full support of policy makers is a prerequisite for carrying out a successful evaluation; impact evaluations should not be undertaken unless policy makers are convinced of the legitimacy of the evaluation and its value for informing important policy decisions.

Third and most important, in a prospective evaluation, the treatment and comparison groups are identified before the intervention being evaluated is implemented. As we will explain in more depth in the chapters that follow, many more options exist for carrying out valid evaluations when the evaluations are planned from the outset before implementation takes place. We argue in parts 2 and 3 that it is almost always possible to find a valid estimate of the counterfactual for any program with clear and transparent assignment rules, provided that the evaluation is designed prospectively. In short, prospective evaluations have the best chance of generating valid counterfactuals. At the design stage, alternative ways to estimate a valid counterfactual can be considered. The design of the impact evaluation can also be fully aligned to program operating rules, as well as to the program's rollout or expansion path.

By contrast, in retrospective evaluations, the team that conducts the evaluation often has such limited information that it is difficult to analyze whether the program was successfully implemented and whether its participants really benefited from it. Many programs do not collect baseline data unless the evaluation has been built in from the beginning, and once the program is in place, it is too late to do so.

Retrospective evaluations using existing data are necessary to assess programs that were established in the past. Options to obtain a valid estimate of the counterfactual are much more limited in those situations. The evaluation is dependent on clear rules of program operation regarding the assignment of benefits. It is also dependent on the availability of data with sufficient coverage of the treatment and comparison groups both before and after program implementation. As a result, the feasibility of a retrospective

evaluation depends on the context and is never guaranteed. Even when feasible, retrospective evaluations often use quasi-experimental methods and rely on stronger assumptions; they thus can produce evidence that is more debatable.[4]

Efficacy Studies and Effectiveness Studies

The main role of impact evaluation is to produce evidence on program performance for the use of government officials, program managers, civil society, and other stakeholders. Impact evaluation results are particularly useful when the conclusions can be applied to a broader population of interest. The question of generalizability is key for policy makers, for it determines whether the results identified in the evaluation can be replicated for groups beyond those studied in the evaluation if the program is scaled up.

In the early days of impact evaluations of development programs, a large share of evidence was based on *efficacy studies:* studies carried out in a specific setting under closely controlled conditions to ensure fidelity between the evaluation design and program implementation. Because efficacy studies are often carried out as pilots with heavy technical involvement from researchers while the program is being implemented, the impacts of these often small-scale efficacy pilots may not necessarily be informative about the impact of a similar project implemented on a larger scale under normal circumstances. Efficacy studies explore proof of concept, often to test the viability of a new program or a specific theory of change. If the program does not generate anticipated impacts under these carefully managed conditions, it is unlikely to work if rolled out under normal circumstances. For instance, a pilot intervention introducing new medical treatment protocols may work in a hospital with excellent managers and medical staff, but the same intervention may not work in an average hospital with less attentive managers and limited staff. In addition, cost-benefit computations will vary, as fixed costs and economies of scale may not be captured in small efficacy studies. As a result, whereas evidence from efficacy studies can be useful to test an innovative approach, the results often have limited generalizability and do not always adequately represent more general settings, which are usually the prime concern of policy makers.

By contrast, *effectiveness studies* provide evidence from interventions that take place in normal circumstances, using regular implementation channels, and aim to produce findings that can be generalized to a large population. When effectiveness evaluations are properly designed and implemented, the results may be generalizable to intended beneficiaries beyond the evaluation sample, so long as the expansion uses the same implementation structures

Key Concept

Efficacy studies assess whether a program *can* work under ideal conditions, while *effectiveness studies* assess whether a programs *does* work under normal conditions.

and reaches similar populations as in the evaluation sample. This external validity is of critical importance to policy makers because it allows them to use the results of the evaluation to inform program-wide decisions that apply to intended beneficiaries beyond the evaluation sample (see box 1.3).

Box 1.3: Testing for the Generalizability of Results: A Multisite Evaluation of the "Graduation" Approach to Alleviate Extreme Poverty

By evaluating a program in multiple contexts, researchers can examine whether the results from an impact evaluation are generalizable. These so-called *multisite evaluations* contribute to the growing body of evidence about what works and what does not in development and can provide important insights for policy makers across countries.

For example, in 2007, Banerjee and others began a multisite evaluation of the "graduation" approach to alleviating extreme poverty. The model had received much attention worldwide after yielding impressive results in Bangladesh. Developed by the Bangladesh Rural Advancement Committee (BRAC), a large global development organization, the model aimed to help "graduate" the very poor from extreme poverty through transfers of cash, productive assets, and intensive training.

Banerjee and his colleagues sought to explore whether the graduation approach would work across countries through six simultaneous randomized impact evaluations in Ethiopia, Ghana, Honduras, India, Pakistan, and Peru. In each country, the researchers worked with local nongovernmental organizations (NGOs) to implement a similar graduation program. While the program was adjusted to fit the different contexts in each country, the key principles remained the same. The program targeted the poorest households in villages in the poorest regions of each country. For 24 months, beneficiary households were given productive assets, training, support, life skills coaching, cash, health information, and help with financial inclusion. The impact evaluation assessed the effectiveness of providing this bundle of benefits.

The study evaluated the impacts of the program on 10 sets of outcomes. One year after the program ended in the six countries, there were significant improvements in 8 out of the 10 sets of outcomes: per capita consumption, food security, asset value, financial inclusion, time spent working, income and revenue, mental health, and political involvement. The magnitude of the impacts varied across countries, with substantial impacts on asset value in all but one country. There were no statistically significant impacts on the physical health index.

The results varied country by country. Improvements in per capita consumption were not significant in Honduras and Peru, and improvements in asset value were not significant in Honduras. In the aggregate, however, the evaluation pointed to the promise of this type of multifaceted intervention in improving the lives of the very poor across a range of settings.

Sources: Banerjee and others 2015; BRAC 2013.

Complementary Approaches

As noted, impact evaluations answer specific cause-and-effect questions. Other approaches—including close *monitoring* of the program, as well as the complementary use of other evaluation approaches such as *ex ante simulations, mixed method analysis* drawing on both qualitative and quantitative data, and *process evaluations*—can serve as valuable complements to impact evaluations. These other approaches have many useful applications, such as to estimate the effect of reforms before they are implemented, to help focus core impact evaluation questions, to track program implementation, and to interpret the results from impact evaluations.

Impact evaluations conducted in isolation from other sources of information are vulnerable in terms of both their technical quality and their policy relevance. While impact evaluation results can provide robust evidence as to whether there has been an effect, they are often limited in providing insights into the channels by which the policy or program affected the observed results. Without information from process evaluations on the nature and content of the program to contextualize evaluation results, policy makers can be left puzzled about why certain results were or were not achieved. Additionally, without monitoring data on how, when, and where the program is being implemented, the evaluation will be blind as to whether and when benefits were received by the intended beneficiaries, or whether benefits reached the comparison group unintentionally.

Monitoring

Monitoring program implementation, most often through the use of administrative data, is critical in an impact evaluation. It lets the evaluation team verify whether activities are being implemented as planned: which participants received the program, how fast the program is expanding, and how resources are being spent. This information is critical to implementing the evaluation, for example, to ensure that baseline data are collected before the program is introduced within the evaluation sample and to verify the integrity of the treatment and comparison groups. Monitoring is critical to checking that a beneficiary actually participates in the program and that a nonbeneficiary does not participate. In addition, administrative data can provide information on the cost of implementing the program, which is also needed for cost-benefit and cost-effectiveness analyses.

Ex Ante Simulations

Ex ante simulations are evaluations that use available data to simulate the expected effects of a program or policy reform on outcomes of interest. They can be very useful in assessing the relative expected effectiveness of a range of alternative program design options on results. These are commonly used methods that depend on the availability of ample high-quality data that can be used to apply simulation models appropriate to the question at hand (see box 1.4). In contrast to impact evaluations, these methods are used to simulate potential future effects, rather than measuring actual impacts of implemented programs. These types of methods can be extremely useful in benchmarking likely program effects and establishing realistic objectives, as well as in estimating costs, rates of return, and other economic parameters. They are often used as the basis for the economic analysis of projects, notably before a reform is introduced or a project is implemented.

Box 1.4: Simulating Possible Project Effects through Structural Modeling: Building a Model to Test Alternative Designs Using Progresa Data in Mexico

A certain type of ex ante simulation—*structural modeling*—can be used to estimate the effects of a program under a range of alternative designs. In the Progresa/Oportunidades/Prospera evaluation described in box 1.1, the data collected were rich enough for researchers to build a model that could simulate expected effects of alternative program designs.

Todd and Wolpin (2006) used baseline data from the impact evaluation to build a model of parental decisions about their children, including child schooling. They simulated what the effects would be under different program designs. They found that if the program eliminated cash incentives for school attendance for lower grades and used the money to increase the cash incentives for students in higher grades, the effects on average schooling completed would likely be larger.

In this case, the projections were done using the baseline survey of an impact evaluation that had been completed. The results of the predictions could be tested to see if they yielded the same impacts as the actual program experiment. This is not generally possible, however. These types of simulation methods are often used before the program is actually implemented to examine the likely effects of various alternative program designs. Thus, they can provide a basis to narrow down the range of options to test in practice.

Source: Todd and Wolpin 2006.
Note: For another example of structural modeling, see Bourguignon, Ferreira, and Leite (2003).

Mixed Methods

Mixed method approaches that combine quantitative and qualitative data are a key supplement to impact evaluations based on the use of quantitative data alone, particularly to help generate hypotheses and focus research questions before quantitative data are collected and to provide perspectives and insights on a program's performance during and after program implementation. There are many qualitative methods, and they comprise their own research domain.[5] Methods generating qualitative data generally employ open-ended approaches that do not rely on predetermined responses from those being interviewed. Data are generated through a range of approaches, including focus groups, life histories, and interviews with selected beneficiaries and other key informants (Rao and Woolcock 2003). They can also include various observational and ethnographic assessments. Although the observations, views, and opinions gathered during qualitative work are usually not statistically representative of the program's beneficiaries—and thus are not generalizable—they are useful to understand why certain results have or have not been achieved (see box 1.5).

Evaluations that integrate qualitative and quantitative analysis are characterized as using *mixed methods* (Bamberger, Rao, and Woolcock 2010).

Box 1.5: A Mixed Method Evaluation in Action: Combining a Randomized Controlled Trial with an Ethnographic Study in India

Mixed methods approaches can be especially helpful when evaluating programs with outcomes that are difficult to measure in quantitative surveys. Programs in democracy and governance are one such example.

When designing an evaluation strategy for the People's Campaign program, which aimed to improve citizen participation in village governments, Ananthpur, Malik, and Rao (2014) integrated a randomized controlled trial (RCT, see glossary) with an ethnographic study conducted in a subset of 10 percent of the evaluation sample used for the RCT. Matching methods were used to ensure similar characteristics between treatment and comparison villages in the sample for the qualitative study. An experienced field investigator was assigned to live in each village and study the impacts of the program on the village social and political structures.

The ethnographic study continued for two years after the RCT ended, allowing for observations of longer-term effects. While the RCT found that the intervention had no statistically significant impact, the qualitative study provided insights into why the intervention failed. The qualitative research identified several factors that hampered the effectiveness of the intervention: variations

(continued)

Box 1.5: A Mixed Method Evaluation in Action: Combining a Randomized Controlled Trial with an Ethnographic Study in India *(continued)*

in the quality of program facilitation, lack of top-down support, and entrenched local power structures.

The qualitative evidence also uncovered some less tangible and unexpected program impacts. In treatment villages, the program improved dispute resolution concerning service delivery and increased women's participation in village development activities. Moreover, the field researchers observed that the village governments functioned better in treatment villages.

Without the nuanced understanding of context and local dynamics provided by the qualitative component, the researchers would not have been able to understand why the quantitative data found no impacts. The ethnographic study was able to provide a richer evaluation, with insights into elements useful to improving the program.

Source: Ananthpur, Malik, and Rao 2014.

In developing a mixed method approach, Creswell (2014) defines three basic approaches:

1. *Convergent parallel.* Both quantitative and qualitative data are collected at the same time and used to triangulate findings or to generate early results about how the program is being implemented and perceived by beneficiaries.
2. *Explanatory sequential.* Qualitative data provide context and explanations for the quantitative results, to explore outlier cases of success and failure, and to develop systematic explanations of the program's performance as it was found in the quantitative results. In this way, qualitative work can help explain why certain results are observed in the quantitative analysis, and can be used to get inside the "black box" of what happened in the program (Bamberger, Rao, and Woolcock 2010).
3. *Exploratory sequential.* The evaluation team can use focus groups, listings, interviews with key informants, and other qualitative approaches to develop hypotheses as to how and why the program would work, and to clarify research questions that need to be addressed in the quantitative impact evaluation work, including the most relevant program design alternatives to be tested through the impact evaluation.

Process Evaluations

Process evaluations focus on how a program is implemented and operates, assessing whether it conforms to its original design and documenting its development and operation. Process evaluations can usually be carried out

relatively quickly and at a reasonable cost. In pilots and in the initial stages of a program, they can be a valuable source of information on how to improve program implementation and are often used as first steps in developing a program so that operational adjustments can be made before the program design is finalized. They can test whether a program is operating as designed and is consistent with the program's theory of change (box 1.6).

Box 1.6: Informing National Scale-Up through a Process Evaluation in Tanzania

There are many facets to a program's performance. Evidence from process evaluations can complement impact evaluation results and provide a more complete picture of program performance. This can be particularly important for pilot programs to shed light on how new institutions and new processes are functioning.

In 2010, the government of Tanzania decided to pilot a community-based conditional cash transfer (CCT) in three districts. The program provided a cash transfer to poor households based on compliance with certain education and health requirements. Community groups assisted in assigning the cash transfer to the most vulnerable households in their communities. To evaluate whether this community-driven system worked in the Tanzanian context, a group of World Bank researchers decided to integrate a process evaluation into a traditional impact evaluation.

The process evaluation used both qualitative and quantitative data. A year after fielding the baseline survey in pilot districts, researchers organized a community scorecard exercise to rate aspects of the program, drawing on focus groups consisting of community members. The focus groups were also used to hold in-depth discussions about program impacts that can be harder to quantify, such as changes in relationships among household members or community dynamics. The aim of the process evaluation was to understand how the program operated in practice and to provide recommendations for improvements.

The impact evaluation found that the program had positive and statistically significant impacts on key education and health outcomes. Children in participant households were about 15 percent more likely to complete primary school and 11 percent less likely to be sick. Focus groups with teachers further revealed that students in treatment groups were more prepared and attentive.

However, focus groups with community members indicated there was a level of discontent with the process of selecting beneficiaries. Participants complained about a lack of transparency in beneficiary selection and delays in payments. The process evaluation allowed program managers to address these issues, improving program operations.

The evaluation work informed the Tanzanian government's decision to scale up the program. The community-based CCT is expected to reach almost 1 million households by 2017, drawing on lessons from this evaluation.

Sources: Berman 2014; Evans and others 2014.

A process evaluation should include the following elements, often drawn from a results chain or logic model (see chapter 2), complemented by program documents and interviews with key informants and beneficiary focus groups:[6]

- Program objectives and the context in which the program is operating
- Description of the process used to design and implement the program
- Description of program operations, including any changes in operations
- Basic data on program operations, including financial and coverage indicators
- Identification and description of intervening events that may have affected implementation and outcomes
- Documentation, such as concept notes, operations manuals, meeting minutes, reports, and memoranda.

Applying an impact evaluation to a program whose operational processes have not been validated poses a risk that either the impact evaluation resources are misspent when a more simple process evaluation would have been sufficient, or that needed adjustments in program design are introduced once the impact evaluation is underway, thereby changing the nature of the program being evaluated and the utility of the impact evaluation.

Cost-Benefit and Cost-Effectiveness Analysis

It is critically important that impact evaluation be complemented with information on the cost of the project, program, or policy being evaluated.

Key Concepts

Cost-benefit analysis estimates the total expected benefits of a program, compared to its total expected costs. Cost-effectiveness analysis compares the relative cost of two or more programs or program alternatives in reaching a common outcome.

Once impact evaluation results are available, they can be combined with information on program costs to answer two additional questions. First, for the basic form of impact evaluation, adding cost information will allow you to perform a cost-benefit analysis, which will answer the question: What is the benefit that a program delivers for a given cost? *Cost-benefit analysis* estimates the total expected benefits of a program, compared to its total expected costs. It seeks to quantify all of the costs and benefits of a program in monetary terms and assesses whether benefits outweigh costs.[7]

In an ideal world, cost analysis based on impact evaluation evidence would exist not only for a particular program, but also for a series of programs or program alternatives, so that policy makers could assess which program or alternative is most cost effective in reaching a particular goal. When an impact evaluation is testing program alternatives, adding cost information allows you to answer the second question: How do various

program implementation alternatives compare in cost-effectiveness? This *cost-effectiveness analysis* compares the relative cost of two or more programs or program alternatives in reaching a common outcome, such as agricultural yields or student test scores.

In a cost-benefit or cost-effectiveness analysis, impact evaluation estimates the benefit or effectiveness side, and cost analysis provides the cost information. This book focuses on impact evaluation and does not discuss in detail how to collect cost data or conduct cost-benefit or cost-effectiveness analysis.[7] However, it is critically important that impact evaluation be complemented with information on the cost of the project, program, or policy being evaluated. Once impact and cost information are available for a variety of programs, cost-effectiveness analysis can identify which investments yield the highest rate of return and allow policy makers to make informed decisions on which intervention to invest in. Box 1.7 illustrates how impact evaluations can be used to identify the most cost-effective programs and improve resource allocation.

Box 1.7: Evaluating Cost-Effectiveness: Comparing Evaluations of Programs That Affect Learning in Primary Schools

By evaluating a number of programs with similar objectives, it is possible to compare the relative cost-effectiveness of different approaches to improving outcomes, such as learning in primary schools. For this to be possible, evaluators must make available not only impact evaluation results, but also detailed cost information on the interventions. In a meta-analysis of learning outcomes in developing countries, Kremer, Brannen, and Glennerster (2013) used cost information from 30 impact evaluations to analyze the cost-effectiveness of different types of education interventions.

The authors compared several types of education interventions, including access to education, business-as-usual inputs, pedagogical innovations, teacher accountability, and school-based management. In particular, they investigated the improvement in test scores, in terms of standard deviations, that could be gained per US$100 spent on the program. Though it is likely that costs would fall if programs were implemented at scale, the researchers used the costs as reported in the evaluations for consistency. They found that pedagogical reforms and interventions that improve accountability and increase teacher incentives tend to be the most cost-effective. On the other hand, the researchers concluded that providing more of the same inputs without changing pedagogy or accountability had limited impacts on test scores. For example, a program in Kenya that increased the number of teachers in schools

(continued)

Box 1.7: Evaluating Cost-Effectiveness: Comparing Evaluations of Programs That Affect Learning in Primary Schools *(continued)*

had no significant impact on test scores for students.

Programs that empowered local communities through school-based management interventions seemed to be the most successful and cost-effective, especially when these reforms were formalized. For instance, while creating and training local school committees in Indonesia did not have significant impacts on test scores, making the committees more representative through elections was highly cost-effective.

As the study by Kremer, Brannen, and Glennerster (2013) illustrates, comparing evaluations of interventions that have similar objectives can shed light on the effectiveness of different interventions across different contexts. Nonetheless, researchers must recognize that contexts vary considerably across programs and settings. It also remains relatively rare to have rich cross-program data with comparable outcome measures, impact evaluations, and cost information.

Source: Kremer, Brannen, and Glennerster 2013.

Ethical Considerations Regarding Impact Evaluation

When the decision is made to design an impact evaluation, some important ethical issues must be considered. Questions have even been raised about whether impact evaluation is ethical in and of itself. One point of departure for this debate is to consider the ethics of investing substantial public resources in programs whose effectiveness is unknown. In this context, the lack of evaluation can itself be seen as unethical. The information on program effectiveness that impact evaluations generate can lead to more effective and ethical investment of public resources.

Other ethical considerations relate to the rules used to assign program benefits, to the methods by which human subjects are studied, and to the transparency in documenting research plans, data, and results. These issues are discussed in detail in chapter 13.

The most basic ethical principle in an evaluation is that the delivery of interventions with known benefits should not be denied or delayed solely for the purpose of the evaluation. In this book, we argue that evaluations should not dictate how benefits are assigned, but that instead evaluations should be fitted to program assignment rules that are equitable and transparent. In this context, any ethical concerns about the rules of program assignment do not stem from the impact evaluation itself but directly from the program operational rules. Planning evaluations can be helpful in

clarifying program operational rules and helping to review whether they are equitable and transparent, based on clear criteria for eligibility.

Randomized assignment of program benefits often raises ethical concerns about denying program benefits to eligible beneficiaries. Yet most programs operate in operational contexts with limited financial and administrative resources, making it impossible to reach all eligible beneficiaries at once. From an ethical standpoint, all subjects who are equally eligible to participate in any type of social program should have the same chance of receiving the program. Randomized assignment fulfills this ethical requirement. In situations where a program will be phased in over time, rollout can be based on randomly selecting the order in which equally deserving beneficiaries will receive the program. In these cases, beneficiaries who enter the program later can be used as a comparison group for earlier beneficiaries, generating a solid evaluation design, as well as a transparent and fair method for allocating scarce resources.

The ethics of impact evaluation go beyond the ethics of program assignment rules. They also include the ethics of conducting research on human subjects, as well as the ethics of conducting transparent, objective, and reproducible research, as explored in chapter 13.

In many countries and international institutions, review boards or ethics committees have been set up to regulate research involving human subjects. These boards are charged with assessing, approving, and monitoring research studies, with the primary goals of protecting the rights and promoting the welfare of all subjects. Although impact evaluations are primarily operational undertakings, they also constitute research studies and as such should adhere to research guidelines for human subjects.

Making your impact evaluation objective, transparent, and reproducible is an equally important ethical component of doing research. To make research transparent, impact evaluation plans can be included in a preanalysis plan and submitted to a study registry. Once the research is completed, the data and code used in the analysis can be made publicly available so that others can replicate the work, while protecting anonymity.

Impact Evaluation for Policy Decisions

Impact evaluations are needed to inform policy makers on a range of decisions, from curtailing inefficient programs, to scaling up interventions that work, to adjusting program benefits, to selecting among various program alternatives. They are most effective when applied selectively to answer important policy questions, and they are often applied to innovative pilot programs that are testing an unproven, but promising approach.

The Mexican conditional cash transfer evaluation described in box 1.1 became influential not only because of the innovative nature of the program, but also because its impact evaluation provided credible and strong evidence that could not be ignored in subsequent policy decisions. The program's adoption and expansion both nationally and internationally were strongly influenced by the evaluation results.

Impact evaluations can be used to explore different types of policy questions. The basic form of impact evaluation will test the effectiveness of a given program. In other words, it will answer the question, is a given program or intervention effective compared to the absence of the program? As discussed in part 2, this type of impact evaluation relies on comparing a treatment group that received the innovation, program, or policy to a comparison group that did not in order to estimate effectiveness. The core challenge in an impact evaluation is to construct a comparison group that is as similar as possible to the treatment group. The degree of comparability between treatment and comparison groups is central to the evaluation's *internal validity* and is therefore fundamental to assessing a program's causal impact.

Impact evaluations are also increasingly being used to test design innovations within a program without a pure comparison group selected from outside of the program. These types of evaluations are often done to see whether a particular design innovation can boost program effectiveness or lower costs (see box 1.8).

Evaluations can also be used to test the effectiveness of program implementation alternatives. For instance, they can answer the following question: When a program can be implemented in several ways, which one is the most effective or cost-effective program modality? In this type of evaluation, two or more approaches or design features within a program can be compared with one another to generate evidence as to which is the most cost-effective alternative for reaching a particular goal. These program alternatives are often referred to as *treatment arms*. For example, a program may wish to test alternative outreach campaigns and select one group to receive a mailing campaign, while another receives house-to-house visits, and yet another receives short message service (SMS) text messages, to assess which is most cost-effective. Impact evaluations testing alternative program treatments normally include one treatment group for each of the treatment arms, as well as a pure comparison group that does not receive any program intervention. These types of evaluations allow decision makers to choose among implementation alternatives, and can be very useful for enhancing program performance and saving costs (box 1.9).

In addition, comparisons can be made among subgroups of recipients within a given evaluation, to answer the following question: Is the program

Box 1.8: Evaluating Innovative Programs: The Behavioural Insights Team in the United Kingdom

Created in 2010 by the British government, the Behavioural Insights Team (BIT) was the first government institution dedicated to improving public services through the application of behavioral science. The objectives of the organization include improving the cost-effectiveness of public services, introducing realistic models of human behavior to policy analysis, and enabling people to make better choices. With this aim, the BIT uses experiments with built-in impact evaluations to test innovative ideas in public policy. Since its creation, the organization has implemented over 150 randomized controlled trials in a wide variety of domestic policy areas, often using administrative data.

The BIT has conducted evaluations of innovations to public services that draw on behavioral science literature. The organization collaborated with a London borough to introduce a lottery incentive to increase voter registration before elections. Residents were randomly assigned to three groups—no lottery, a lottery with a prize of £1,000 if they registered before a certain date, and a lottery with a prize of £5,000 if they registered before the same date. The BIT found that the lottery incentive significantly increased voter registration. Moreover, it saved the local government a lot of money; the government had previously relied on an expensive door-to-door canvas to increase voter registration.

In another innovative evaluation, the BIT partnered with the National Health Service and the Department of Health to examine how to cost-effectively encourage people to register as organ donors. This was one of the largest randomized controlled trials ever in the U.K. public sector. The researchers found encouraging results from an intervention that tested the use of different messages on a high traffic government webpage. The best performing short phrase was based on the idea of reciprocity and asked, if you needed an organ transplant, would you have one? If so, help others.

The BIT is jointly owned and financed by the British government; the innovation charity, Nesta; and the employees themselves. The model has spread outside of the United Kingdom, with BIT offices created in Australia and the United States. Moreover, the United States followed the BIT model to establish a Social and Behavioral Science Initiative in the White House in 2015.

Source: Behavioural Insights Team, http://www.behaviouralinsights.co.uk.

more effective for one subgroup than it is compared with another subgroup? For example, did the introduction of a new curriculum raise test scores more among female students than male students? This type of impact evaluation questions seeks to document whether there is some heterogeneity in program impacts across subgroups. Such questions need to be considered upfront, as they need to be incorporated into the design of an impact evaluation and require sufficiently large samples to carry out the analysis of the different subgroups of interest.

Box 1.9: Evaluating Program Design Alternatives: Malnourishment and Cognitive Development in Colombia

In the early 1970s, the Human Ecology Research Station, in collaboration with the Colombian ministry of education, implemented a pilot program to address childhood malnutrition in Cali, Colombia, by providing health care and educational activities, as well as food and nutritional supplements. As part of the pilot, a team of evaluators was tasked to determine how long such a program should last to reduce malnutrition among preschool children from low-income families, and whether the interventions could also lead to improvements in cognitive development.

The program was eventually made available to all eligible families, but during the pilot, the evaluators were able to compare similar groups of children who received different durations of treatment. The evaluators first used a screening process to identify a target group of 333 malnourished children. These children were then classified into 20 sectors by neighborhood, and each sector was randomly assigned to one of four treatment groups. The groups differed only in the sequence in which they started the treatment, and thus in the amount of time that they spent in the program. Group 4 started the earliest and was exposed to the treatment for the longest period, followed by groups 3, 2, and then 1. The treatment itself consisted of six hours of health care and educational activities per day, plus additional food and nutritional supplements. At regular intervals over the course of the program, the evaluators used cognitive tests to track the progress of children in all four groups.

The evaluators found that the children who were in the program for the longest time demonstrated the greatest gains in cognitive improvement. On the Stanford-Binet intelligence test, which estimates mental age minus chronological age, group 4 children averaged –5 months, and group 1 children averaged –15 months.

This example illustrates how program implementers and policy makers are able to use evaluations of multiple treatment arms to determine the most effective program alternative.

Source: McKay and others 1978.

Beyond the various design features already discussed, it is useful to consider the channels through which impact evaluations affect policy. This can happen within a program with respect to decisions about continuing, reforming, or ending a program. Impact evaluation results can also inform the scale-up of pilots, as the Mozambique case in box 1.2 illustrates.

Evaluations can also bring evidence from one country to another or can be used to explore fundamental questions such as those concerning behavior. Venturing beyond the borders of an individual program evaluation raises the question of generalizability. As chapter 4 discusses in

the context of a particular evaluation, the evaluation sample is designed to be statistically representative of the population of eligible units from which the evaluation sample is drawn, and thus externally valid. Beyond external validity, generalizability concerns whether results from an evaluation carried out locally will hold true in other settings and among other population groups. This more expansive and ambitious concept depends on the accumulation of credible empirical evidence across a range of settings.

Increasingly, the impact evaluation field is seeking to build on the growing stock of credible evaluations to achieve broadly generalizable findings. This effort centers on testing whether a particular theory of change holds in different contexts and on exploring whether a similar program tested in different settings yields similar results (see box 1.10). The use of multiple evaluations to answer core questions or assemble evidence through meta-analyses, systematic reviews, and evaluation registries is growing rapidly and opening a new frontier in evaluation work. If results are consistent across multiple settings, this gives policy makers greater confidence in the viability of the program across a range of contexts and population groups. This is an important consideration, as debates about the ability to replicate results are fundamental to questions about the broader effectiveness and scalability of a particular program.

Box 1.10: The Impact Evaluation Cluster Approach: Strategically Building Evidence to Fill Knowledge Gaps

Although the generalizability of a single impact evaluation may be low, in combination with similar evaluations across different contexts, development practitioners can develop more broadly applicable conclusions about what works and what does not. Increasingly, impact evaluation initiatives such as the World Bank's Strategic Impact Evaluation Fund (SIEF) and Development Impact Evaluation (DIME), as well as the International Initiative for Impact Evaluation (3IE), aim to provide policy makers with insights into how program and policy interventions can be more broadly applied, using a *research cluster* approach.

Often calls for proposals are oriented around a set of research questions aimed to inform program and policy design, to generate impact evaluations that will contribute to a coordinated evidence base. The objective is to orient research and the generation of evidence around types of interventions or types of outcomes.

Within these *clusters,* evaluations are generated to fill gaps in the existing body of evidence. For example, there is solid evidence

(continued)

Box 1.10: The Impact Evaluation Cluster Approach: Strategically Building Evidence to Fill Knowledge Gaps *(continued)*

showing that children who receive a combination of nutrition, cognitive stimulation, and health support in the first 1,000 days of life are more likely to avoid developmental delays. However, there is a lack of research on how to best deliver this combined support in scalable and cost-effective ways. SIEF is supporting research to explore this question in Bangladesh, Colombia, India, Indonesia, Madagascar, Mozambique, Nepal, and Niger.

Clustering evaluations around a common set of research questions and using a core set of metrics to measure outcomes helps policy makers and development practitioners see which types of programs work in multiple settings. They can then review their own policy and program designs with a better sense of the contexts in which particular programs have worked or not worked or with respect to how particular outcomes have been achieved across several cases.

Sources: DIME (http://www.worldbank.org/dime); SIEF (http://www.worldbank.org/en/programs/sief-trust-fund); 3IE (http://www.3ieimpact.org).

Deciding Whether to Carry Out an Impact Evaluation

Not all programs warrant an impact evaluation. Impact evaluations should be used selectively when the question being posed calls for a strong examination of causality. Impact evaluations can be costly if you collect your own data, and your evaluation budget should be used strategically. If you are starting, or thinking about expanding, a new program and wondering whether to go ahead with an impact evaluation, asking a few basic questions will help with the decision.

The first question to ask is, what is at stake? Will evidence about the success of the program, program modality, or design innovation inform important decisions? These decisions often involve budgetary allocations and program scale. If there are limited budget implications or if the results will affect only a few people, it may not be worth doing an impact evaluation. For example, it may not be worth conducting an impact evaluation of a program in a small clinic that provides counseling to hospital patients using volunteers. By contrast, a pay reform for teachers that will eventually affect all primary teachers in the country would be a program with much higher stakes.

If you determine that the stakes are high, then the next question is, does any evidence exist to show that the program works? In particular, do you

know how big the program's impact would be? Is there evidence available from similar programs under similar circumstances? If no evidence is available about the potential of the type of program being contemplated, you may want to start out with a pilot that incorporates an impact evaluation. By contrast, if evidence is available from similar circumstances, the cost of an impact evaluation will probably be justified only if it can address an important and new policy question. That would be the case if your program includes some important innovations that have not yet been tested.

To justify mobilizing the technical and financial resources needed to carry out a high-quality impact evaluation, the intervention to be evaluated should be:

- *Innovative*. It will test a new, promising approach.
- *Replicable*. It can be scaled up or can be applied in a different setting.
- *Strategically relevant*. The evidence provided by the impact evaluation will inform an important decision concerning the intervention. This could relate to program expansion, reform, or budget allocations.
- *Untested*. Little is known about the effectiveness of the program or design alternatives, globally or in a particular context.
- *Influential*. The results will be used to inform policy decisions.

A final question to ask is, do we have the resources necessary for a good impact evaluation? These resources concern technical elements such as appropriate data and time, financial resources to carry out the evaluation, as well as institutional resources with respect to the teams involved and their interest in and commitment to building and using causal evidence. As discussed in more depth in chapter 12, an evaluation team is essentially a partnership between two groups: a team of policy makers and a team of researchers. The teams need to work toward the common goal of ensuring that a well-designed, technically robust evaluation is implemented properly and delivers results relevant to key policy and program design questions. A clear understanding of the premise and the promise of impact evaluation by the evaluation team will help ensure its success.

If you decide that an impact evaluation makes sense given the questions at hand and the related need to examine causality, the stakes associated with the results, and the need for evidence about your program's performance, then keep reading—this book is for you and your evaluation team.

Additional Resources

- For accompanying material to this chapter and hyperlinks to additional resources, please see the Impact Evaluation in Practice website (www.worldbank.org/ieinpractice).
- For additional information on impact evaluations, see Khandker, Shahidur R., Gayatri B. Koolwal, and Hussain Samad. 2009. *Handbook on Quantitative Methods of Program Evaluation*. Washington, DC: World Bank.
- For a good overview of randomized controlled trials, see Glennerster, Rachel, and Kudzai Takavarasha. 2013. *Running Randomized Evaluations: A Practical Guide*. Princeton, NJ: Princeton University Press.
- Other resources on randomized controlled trials include the following:
  - Duflo, E., R. Glennerster, and M. Kremer. 2007. "Using Randomization in Development Economics Research: A Toolkit." In *Handbook of Development Economics*, volume 4, edited by T. Paul Schultz and John Strauss, 3895–962. Amsterdam: Elsevier.
  - Duflo, Esther, and Michael Kremer. 2008. "Use of Randomization in the Evaluation of Development Effectiveness." In Vol. 7 of *Evaluating Development Effectiveness*. Washington, DC: World Bank.
- Other useful impact evaluation resources include the following:
  - Leeuw, Frans, and Jos Vaessen. 2009. *Impact Evaluations and Development: NONIE Guidance on Impact Evaluation*. Washington, DC: NONIE.
  - Ravallion, Martin. 2001. "The Mystery of the Vanishing Benefits: Ms. Speedy Analyst's Introduction to Evaluation." *World Bank Economic Review* 15 (1): 115–40.
  - ——. 2007. "Evaluating Anti-Poverty Programs." In Vol. 4 of *Handbook of Development Economics,* edited by T. Paul Schultz and John Strauss. Amsterdam: North Holland.
  - ——. 2009. "Evaluation in the Practice of Development." *World Bank Research Observer* 24 (1): 29–53.

Notes

1. For an overview of conditional cash transfer programs and the influential role played by Mexico's program and its impact evaluation, see Fiszbein and Schady (2009).
2. Administrative data are those data routinely collected as part of program administration and include information on costs, registration and transactions, usually as part of service delivery.
3. There many typologies for evaluations and evaluation questions. See Berk and Rossi (1998) and Rossi, Lipsey, and Freeman (2003)
4. *Quasi-experimental* methods are impact evaluation methods that use a counterfactual but are distinct from *experimental* methods in that quasi-experimental methods are *not* based on randomized assignment of the intervention. See section 2 for a discussion of both types of methods.

5. For an overview of qualitative research methods, see Patton (1990).
6. Adapted from the Bureau of Justice Assistance (1997, 97–98 and 102–3).
7. For a detailed discussion of cost-benefit analysis, see Zerbe and Dively (1994); Brent (1996); Belli and others (2001); and Boardman and others (2001).

References

Ananthpur, Kripa, Kabir Malik, and Vijayendra Rao. 2014. "The Anatomy of Failure: An Ethnography of a Randomized Trial to Deepen Democracy in Rural India." Policy Research Working Paper 6958, World Bank, Washington, DC.

Bamberger, Michael, Vijayendra Rao, and Michael Woolcock. 2010. "Using Mixed Methods in Monitoring and Evaluation: Experiences from International Development." Policy Research Working Paper 5245, World Bank, Washington, DC.

Banerjee, Abhijit, Esther Duflo, Nathanael Goldberg, Dean Karlan, Robert Osei, and others. 2015. "A Multifaceted Program Causes Lasting Progress for the Very Poor: Evidence from Six Countries." *Science* 348 (6236). doi:10.1126/science.1260799.

Behrman, Jere R., and John Hoddinott. 2001. "An Evaluation of the Impact of PROGRESA on Pre-school Child Height." FCND Briefs 104, International Food Policy Research Institute, Washington, DC.

Belli, Pedro, Jock Anderson, Howard Barnum, John Dixon, and Jee-Peng Tan. 2001. *Handbook of Economic Analysis of Investment Operations*. Washington, DC: World Bank.

Berk, Richard A., and Peter Rossi. 1998. *Thinking about Program Evaluation,* second edition. Thousand Oaks, CA: Sage Publications.

Berman, Daphna. 2014. "Tanzania: Can Local Communities Successfully Run Cash Transfer Programs?" Human Development Network, World Bank, Washington, DC.

Boardman, Anthony, Aidan Vining, David Greenberg, and David Weimer. 2001. *Cost-Benefit Analysis: Concepts and Practice.* New Jersey: Prentice Hall.

Bourguignon, François, Francisco H. G. Ferreira, and Phillippe G. Leite. 2003. "Conditional Cash Transfers, Schooling, and Child Labor: Micro-Simulating Brazil's Bolsa Escola Program." *The World Bank Economic Review* 17 (2): 229–54.

BRAC (Bangladesh Rural Advancement Committee). 2013. "An End in Sight for Ultra-poverty." BRAC Briefing Note, November. http://www.brac.net/sites/default/files/BRAC%20Briefing%20-%20TUP.pdf.

Brent, Robert. 1996. *Applied Cost-Benefit Analysis.* Cheltenham, U.K.: Edward Elgar.

Bureau of Justice Assistance. 1997. *Urban Street Gang Enforcement*. Report prepared by the Institute for Law and Justice, Inc. Washington, DC: Office of Justice Programs, Bureau of Justice Assistance, U.S. Department of Justice.

Creswell, John W. 2014. *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches*. Thousand Oaks, CA: SAGE Publications.

Evans, David K., Stephanie Hausladen, Katrina Kosec, and Natasha Reese. 2014. "Community-based Conditional Cash Transfers in Tanzania: Results from a Randomized Trial." World Bank, Washington, DC.

Fiszbein, Ariel, and Norbert Schady. 2009. *Conditional Cash Transfers, Reducing Present and Future Poverty.* Policy Research Report 47603. Washington, DC: World Bank.

Gertler, Paul J. 2004. "Do Conditional Cash Transfers Improve Child Health? Evidence from PROGRESA's Control Randomized Experiment." *American Economic Review* 94 (2): 336–41.

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

Imas, Linda G. M., and Ray C. Rist. 2009. *The Road to Results: Designing and Conducting Effective Development Evaluations*. Washington, DC: World Bank.

Kremer, Michael, Conner Brannen, and Rachel Glennerster. 2013. "The Challenge of Education and Learning in the Developing World." *Science* 340 (6130): 297–300.

Khandker, Shahidur, Gayatri B. Koolwal, and Hussain A. Samad. 2010. *Handbook on Impact Evaluation: Quantitative Methods and Practices.* Washington, DC: World Bank.

Levy, Santiago, and Evelyne Rodríguez. 2005. *Sin Herencia de Pobreza: El Programa Progresa-Oportunidades de México.* Washington, DC: Inter-American Development Bank.

Martinez, Sebastian, Sophie Nadeau, and Vitor Pereira, 2012. "The Promise of Preschool in Africa: A Randomized Impact Evaluation of Early Childhood Development in Rural Mozambique." Washington, DC: World Bank and Save the Children.

McKay, Harrison, Arlene McKay, Leonardo Siniestra, Hernando Gomez, and Pascuala Lloreda. 1978. "Improving Cognitive Ability in Chronically Deprived Children." *Science* 200 (21): 270–78.

Patton, M. Q. 1990. *Qualitative Evaluation and Research Methods*, second edition. Newbury Park, CA: Sage.

Rao, Vijayendra, and Michael Woolcock. 2003. "Integrating Qualitative and Quantitative Approaches in Program Evaluation." In *The Impact of Economic Policies on Poverty and Income Distribution: Evaluation Techniques and Tools*, edited by F. J. Bourguignon and L. Pereira da Silva, 165–90. New York: Oxford University Press.

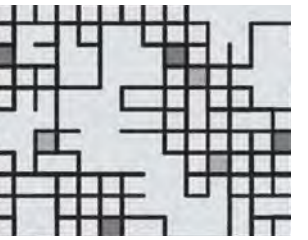
Rossi, Peter, Mark W. Lipsey, and Howard Freeman. 2003. *Evaluation: A Systematic Approach,* seventh edition. Thousand Oaks, CA: Sage Publications.

Schultz, Paul. 2004. "School Subsidies for the Poor: Evaluating the Mexican Progresa Poverty Program." *Journal of Development Economics* 74 (1): 199–250.

Skoufias, Emmanuel, and Bonnie McClafferty. 2001. "Is Progresa Working? Summary of the Results of an Evaluation by IFPRI." International Food Policy Research Institute, Washington, DC.

Todd, Petra, and Kenneth Wolpin. 2006. "Using Experimental Data to Validate a Dynamic Behavioral Model of Child Schooling and Fertility: Assessing the Impact of a School Subsidy Program in Mexico." *American Economic Review* 96 (5): 1384–417.

Zerbe, Richard, and Dwight Dively. 1994. *Benefit Cost Analysis in Theory and Practice.* New York: Harper Collins Publishing.



CHAPTER 2

Preparing for an Evaluation

Initial Steps

This chapter reviews the initial steps in setting up an evaluation. The steps include constructing a theory of change that outlines how the project is supposed to achieve the intended results, developing a results chain as a useful tool for outlining the theory of change, specifying the evaluation question(s), and selecting indicators to assess performance.

These steps are necessary to prepare for an evaluation. They are best taken at the outset of the program or reform being evaluated, when it is first being designed. The steps involve engaging a range of stakeholders—from policy makers to program implementers—to forge a common vision of the program's goals and how they will be achieved. This engagement builds consensus regarding the focus of the evaluation and the main questions to be answered, and will strengthen links between the evaluation, program implementation, and policy. Applying the steps lends clarity and specificity that are useful both for developing a good impact evaluation and for designing and implementing an effective program. Each step is clearly defined and articulated within the logic model embodied in the results chain—from a precise specification of goals and questions, to the articulation of ideas embodied in the theory of change, to the identification of the outcomes the program aims to provide. A clear specification of the particular indicators that will be

used to measure program success is needed not only to ensure that the evaluation is focused, but also that the program has well-defined objectives. It also provides a basis for determining anticipated effect sizes from the program. These parameters are essential to establishing technical elements of the evaluation, including the size of the sample required for the evaluation and power calculations, as reviewed in chapter 15.

In most impact evaluations, it will be important to include an assessment of cost-benefit or cost-effectiveness, as discussed in chapter 1. Policy makers are always concerned with learning not only which programs or reforms are effective, but also at what cost. This is a crucial consideration for informing decisions about whether a program could be scaled up and replicated—a concern that is central to policy decisions.

Constructing a Theory of Change

A *theory of change* is a description of how an intervention is supposed to deliver the desired results. It describes the causal logic of how and why a particular program, program modality, or design innovation will reach its intended outcomes. A theory of change is a key underpinning of any impact evaluation, given the cause-and-effect focus of the research. As one of the first steps in the evaluation design, constructing a theory of change can help specify the research questions.

Theories of change depict a sequence of events leading to outcomes; they explore the conditions and assumptions needed for the change to take place, make explicit the causal logic behind the program, and map the program interventions along logical causal pathways. Working with the program's stakeholders to put together a theory of change can clarify and improve program design. This is especially important in programs that seek to influence behavior: theories of change can help disentangle the intervention's inputs and activities, the outputs that are delivered, and the outcomes that stem from expected behavioral changes among beneficiaries.

The best time to develop a theory of change for a program is at the beginning of the design process, when stakeholders can be brought together to develop a common vision for the program, its goals, and the path to achieving those goals. Stakeholders can then start implementing the program from a common understanding of the program, its objectives, and how it works.

Program designers should also review the literature for accounts of experience with similar programs, and verify the contexts and assumptions behind the causal pathways in the theory of change they are outlining. For example, in the case of the project in Mexico (described in box 2.1) that replaced dirt floors with cement floors, the literature provided valuable information on how parasites are transmitted and how parasite infestation leads to childhood diarrhea.

Box 2.1: Articulating a Theory of Change: From Cement Floors to Happiness in Mexico

In their evaluation of the Piso Firme or "firm floor" project, Cattaneo and others (2009) examined the impact of housing improvements on health and welfare. Both the project and the evaluation were motivated by a clear theory of change.

The objective of the Piso Firme project is to improve the living standards—especially the health—of vulnerable groups living in densely populated, low-income areas of Mexico. The program was first started in the northern State of Coahuila and was based on a situational assessment conducted by the state government.

The program's results chain is clear. Eligible neighborhoods are surveyed door to door, and households are offered up to 50 square meters of cement. The government purchases and delivers the cement, and the households and community volunteers supply the labor to install the floor. The output is the construction of a cement floor, which can be completed in about a day. The expected outcomes of the improved home environment include cleanliness, health, and happiness.

The rationale for this results chain is that dirt floors are a vector for parasites because they are harder to keep clean. Parasites live and breed in feces and can be ingested by humans when they are tracked into the home by animals or people. Evidence shows that young children who live in houses with dirt floors are more likely to be infected with intestinal parasites, which can cause diarrhea and malnutrition, often leading to impaired cognitive development or even death. Cement floors interrupt the transmission of parasitic infestations. They also control temperature better and are more aesthetically pleasing.

Those expected outcomes informed the research questions that Cattaneo and others (2009) addressed in the evaluation. They hypothesized that replacing dirt floors with cement floors would reduce the incidence of diarrhea, malnutrition, and micronutrient deficiency. In turn, improved health and nutrition should result in improved cognitive development in young children. The researchers also anticipated and tested for improvements in adult welfare, as measured by people's increased satisfaction with their housing situation and lower rates of depression and perceived stress.

Source: Cattaneo and others 2009.

Developing a Results Chain

A results chain is one way of depicting a theory of change. Other approaches include theoretical models, logic models, logical frameworks, and outcome models. Each of these models includes the basic elements of a theory of change: a causal chain, a specification of outside conditions and influences, and key assumptions. In this book, we will use the results chain model because we find that it is the simplest and clearest model to outline the theory of change in the operational context of development programs.

Key Concept

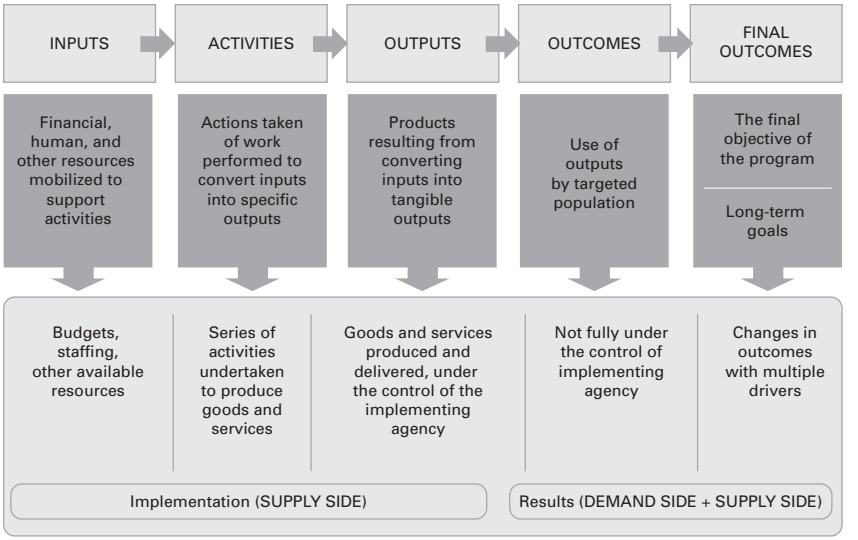
A results chain sets out the sequence of inputs, activities, and outputs that are expected to improve outcomes and final outcomes.

A results chain establishes the causal logic from the initiation of the program, beginning with resources available, to the end, looking at long-term goals. It sets out a logical, plausible outline of how a sequence of inputs, activities, and outputs for which a program is directly responsible interacts with behavior to establish pathways through which impacts are achieved (figure 2.1). A basic results chain will map the following elements:

- *Inputs.* Resources at the disposal of the project, including staff and budget.
- *Activities.* Actions taken or work performed to convert inputs into outputs.
- *Outputs.* The tangible goods and services that the project activities produce; these are directly under the control of the implementing agency.
- *Outcomes.* Results likely to be achieved once the beneficiary population uses the project outputs; these are usually achieved in the short to medium term and are usually *not* directly under the control of the implementing agency.
- *Final outcomes.* The final results achieved indicating whether project goals were met. Typically, final outcomes can be influenced by multiple factors and are achieved over a longer period of time.

The results chain covers both implementation and results. *Implementation* concerns the work delivered by the project, including inputs, activities, and outputs. These are the areas under the direct responsibility of the project that are usually monitored to verify whether the project is delivering the goods and services as intended. *Results* consist of the outcomes and final outcomes, which are not under the direct control of the project and which are contingent on behavioral changes by program beneficiaries. In other words, they depend on the interactions between

Figure 2.1 The Elements of a Results Chain



the supply side (implementation) and the demand side (beneficiaries). These are the areas typically subject to impact evaluation to measure effectiveness.

A good results chain will help surface assumptions and risks implicit in the theory of change. Policy makers are best placed to articulate the causal logic and the assumptions on which it relies—as well as the risks that may affect the achievement of intended results. The team that conducts the evaluation should draw out these implicit assumptions and risks in consultation with policy makers. A good results chain will also reference evidence from the literature regarding the performance of similar programs.

Results chains are useful for all projects—regardless of whether or not they will include an impact evaluation—because they allow policy makers and program managers to make program goals explicit, thus helping to clarify the causal logic and sequence of events behind a program. They can also identify gaps and weak links in program design and therefore can help improve program design. Results chains also facilitate monitoring and evaluation by making evident what information needs to be monitored along each link in the chain to track program implementation and what outcome indicators need to be included when the project is evaluated.

Specifying Evaluation Questions

A clear evaluation question is the starting point of any effective evaluation. The formulation of an evaluation question focuses the research to ensure that it is tailored to the policy interest at hand. In the case of an impact evaluation, it needs to be structured as a testable hypothesis. The impact evaluation then generates credible evidence to answer that question. As you will remember, the basic impact evaluation question is, what is the impact (or causal effect) of a program on an outcome of interest? The focus is on the *impact*: that is, the changes *directly attributable* to a program, program modality, or design innovation.

The evaluation question needs to be guided by the core policy interest at hand. As discussed in chapter 1, impact evaluations can explore a range of questions. The evaluation team should clarify which question will be examined as a first step, drawing on the theory of change before looking at how the evaluation will be conducted.

Traditionally, impact evaluations have focused on the impact of a fully implemented program on final outcomes, compared with the results observed in a comparison group that did not benefit from the program. The use of impact evaluations is expanding. The evaluation team can ask, is the key evaluation question the "classic" question about the effectiveness of a program in changing final outcomes? Or is it about testing whether one program modality is more cost effective than another? Or is it about introducing a program design innovation that is expected to change behaviors, such as enrollment? New approaches to impact evaluation are being introduced in creative ways to tackle questions of policy concern across a range of disciplines (see box 2.2).

In an impact evaluation, the evaluation question needs to be framed as a *well-defined, testable hypothesis.* You need to be able to frame the question in such a way that you can quantify the difference between the results obtained contrasting the treatment and comparison groups. The results chain can be used as a basis for formulating the hypothesis that you would like to test using the impact evaluation. As illustrated in box 2.3, there are often a few hypotheses associated with the program, but not all can or should be explored in an impact evaluation. In the mathematics curriculum example in box 2.2, the evaluation question was derived from fundamental elements of the theory of change and formulated as a clear, testable, and quantifiable hypothesis: What is the effect of a new mathematics curriculum on test scores? In the example that we will apply throughout the book, the Health Insurance Subsidy Program (HISP), the evaluation question is, what is the effect of HISP on poor households' out-of-pocket health expenditures?

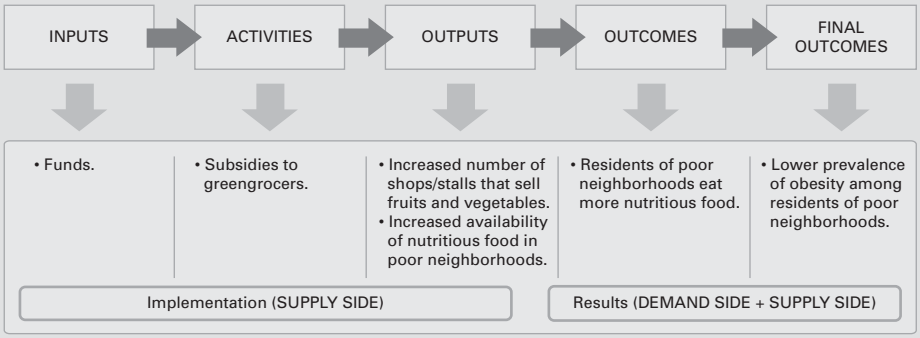
Box 2.2: Mechanism Experiments

A *mechanism experiment* is an impact evaluation that tests a particular causal mechanism within the theory of change. Say you have identified an issue and a possible program to remedy the issue. You are thinking of designing an evaluation to test the effectiveness of the program. Should your evaluation directly test the impact of the program? A recent stream of thought argues that such a program evaluation might not always be the best way to start out, and that in some cases it might be preferable not to carry out a program evaluation but rather to test some of the underlying assumptions or mechanisms. Mechanism experiments do not test a program: they test a causal mechanism that underlies the choice of a program.

For example, you might be worried that people living in poor neighborhoods of a city have higher rates of obesity than people living in more affluent parts of the same city. After some research, you observe that poor neighborhoods have fewer shops and stalls that sell fresh fruits and vegetables and other nutritious food. You think that this lack of supply may be contributing to obesity, and that you may be able to remedy the situation by offering subsidies to greengrocers to set up more sales points. A simple results chain is outlined below (see figure B2.2.1).

A program evaluation would focus on testing the impact of subsidies to greengrocers in a set of poor neighborhoods. By contrast, a mechanism experiment might more directly test your underlying assumptions. For example, it might test the following assumption: If residents of poor neighborhoods have more access to nutritious food, they will eat more of it. One way of testing this would be to distribute a free weekly basket of fruits and vegetables to a group of residents and compare their intake of fruits and vegetables to that of residents who do not receive the free basket. If you find no differences in fruit and vegetable intakes in this mechanism experiment, it is unlikely that providing subsidies to greengrocers would have a significant impact

Figure B2.2.1 Identifying a Mechanism Experiment from a Longer Results Chain



(continued)

Box 2.2: Mechanism Experiments *(continued)*

either, because one of the underlying causal mechanisms is not working.

A mechanism experiment should normally be much cheaper to implement than a full program evaluation because you can carry it out at a smaller scale. In the obesity example, it would be quite expensive to provide subsidies to greengrocers in many neighborhoods and survey a large number of residents in those neighborhoods. By contrast, the free grocery baskets would be much cheaper, and it would be sufficient to enroll a few hundred families. If the mechanism experiment shows that the mechanism works, you would still need to carry out a policy experiment to assess whether the subsidies are an effective way of making fruits and vegetables available to residents of poor neighborhoods.

Source: Ludwig, Kling, and Mullainathan 2011.

Box 2.3: A High School Mathematics Reform: Formulating a Results Chains and Evaluation Question

Imagine that the ministry of education of country A is thinking of introducing a new high school mathematics curriculum. This curriculum is designed to be more intuitive for teachers and students, improve students' performance on standardized mathematics tests, and ultimately, improve students' ability to complete high school and access better jobs. The following results chain outlines the theory of change for the program (see figure B2.3.1).

- The inputs include staff from the ministry of education to spearhead the reform, high school mathematics teachers, a budget to develop the new math curriculum, and the municipal facilities where the mathematics teachers will be trained.
- The program's activities consist of designing the new mathematics curriculum; developing a teacher training program; training the teachers; and commissioning, printing, and distributing new textbooks.
- The outputs are the number of teachers trained, the number of textbooks delivered to classrooms, and the adaptation of standardized tests to the new curriculum.
- The short-term outcomes consist of teachers' use of the new methods and textbooks in their classrooms and their application of the new tests.
- The medium-term outcomes are improvements in student performance on the standardized mathematics tests.
- Final outcomes are increased high school completion rates and higher employment rates and earnings for graduates.

Several hypotheses underlie the theory of change:

- Trained teachers use the new curriculum effectively.

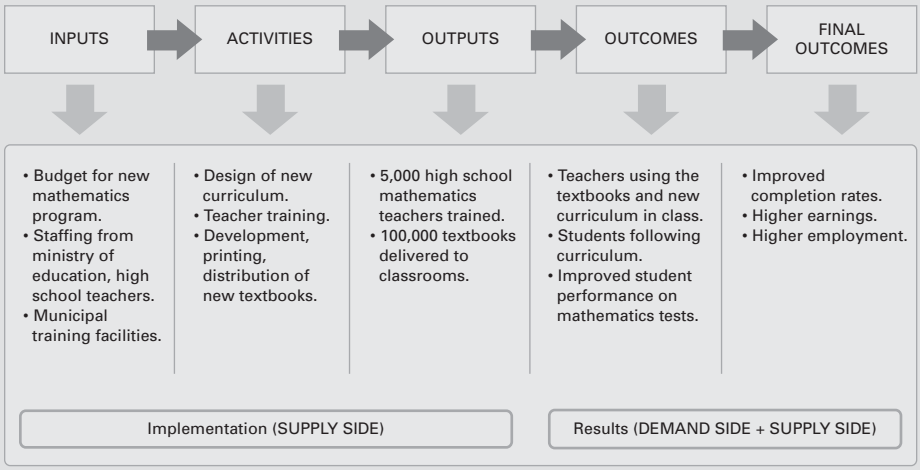
(continued)

Box 2.3: A High School Mathematics Reform: Formulating a Results Chains and Evaluation Question *(continued)*

- If the teachers are trained and the textbooks are distributed, these will be applied and the students will follow the curriculum.
- The new curriculum is superior to the old one in imparting knowledge of mathematics.
- If implementation is carried out as planned, then the math test results will improve by 5 points, on average.
- Performance in high school mathematics influences high school completion rates, employment prospects, and earnings.

The core evaluation question developed by the evaluation team of policy makers in the ministry of education and the researchers engaged in determining the effectiveness of the program is, what is the effect of the new mathematics curriculum on test scores? This question goes to the heart of the policy interest concerning the effectiveness of the new curriculum.

Figure B2.3.1 A Results Chain for the High School Mathematics Curriculum Reform



The Health Insurance Subsidy Program (HISP): An Introduction

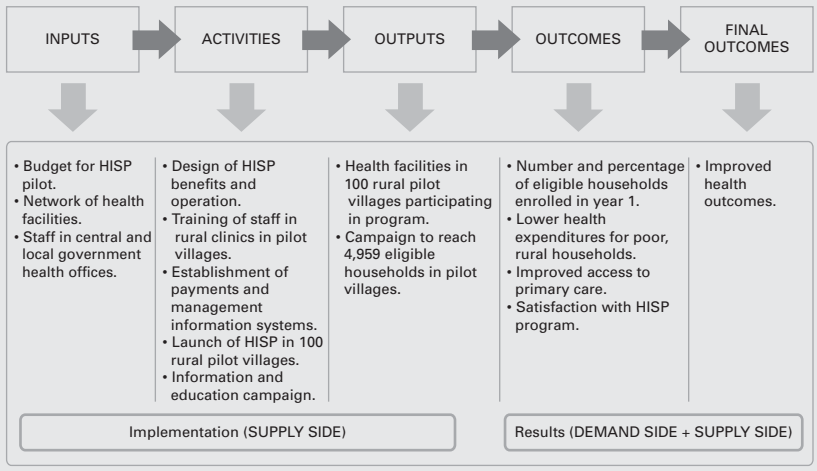
The Health Insurance Subsidy Program (HISP) is a fictional case of a government undertaking large-scale health sector reform. Questions related to this case will be used throughout the book. The Impact Evaluation in Practice website (www.worldbank.org/ieinpractice) contains solutions for the HISP case study questions, a data set, and the analysis code in Stata, as well as an online technical companion that provides a more formal treatment of data analysis.

The ultimate objective of HISP is improving the health of the country's population. The innovative—and potentially costly—HISP is being piloted. The government is concerned that poor rural households are unable to afford the costs of basic health care, with detrimental consequences for their health. To address this issue, HISP subsidizes health insurance for poor rural households, covering costs related to primary health care and medicine. The central objective of HISP is to reduce the cost of health care for poor families and, ultimately, to improve health outcomes. Policy makers are considering expanding HISP to cover the whole country, which would cost hundreds of millions of dollars.

The results chain for HISP is illustrated in figure 2.2. The hypotheses related to the HISP reform assume the following: that households will enroll in the program once it is offered, that enrollment in the program will lower households' out-of-pocket health expenditures, that costs are preventing rural populations from accessing available health care and medicine, and that out-of-pocket expenditures on health-related costs are a core contributor to poverty and poor health outcomes.

The key evaluation question is this: What is the impact of HISP on poor households' out-of-pocket health expenditures? Throughout the book and in the online material, we will answer this same evaluation question with regard to HISP several times, using different methodological approaches. You will see that different—and sometimes conflicting—answers will emerge, depending on what evaluation methodology is used.

Figure 2.2 The HISP Results Chain



Selecting Outcome and Performance Indicators

A clear evaluation question needs to be accompanied by the specification of which outcome measures will be used to assess results, including in the case of multiple outcomes. The outcome measures selected will be used to determine whether or not a given program or reform is successful. They are also the indicators that can be referenced in applying power calculations used to determine the sample sizes needed for the evaluation, as discussed in chapter 15.

Once the main indicators of interest are selected, clear objectives in terms of program success need to be established. This step amounts to determining the anticipated effect of the program on the core outcome indicators that have been selected. *Effect sizes* are the changes expected as a result of the program or the reform, such as the change in test scores or the take-up rate of a new type of insurance policy. Expected effect sizes are the basis for conducting power calculations.

It is critical to have the main stakeholders in the evaluation team (both the research team and the policy team) agree on both the primary outcome indicators of interest in the impact evaluation and the effect sizes anticipated as a result of the program or innovation (for more on the evaluation team, see chapter 12). These are the indicators that will be used to judge program success and form the basis for the power calculations. Impact evaluations can fail because they do not have sample sizes large enough to detect the changes that have resulted from the program; they are "underpowered." Minimum expected effect sizes should be determined to establish basic criteria for success of the program or innovation. When data are available, ex ante simulations can be conducted to look at different outcome scenarios to benchmark the type of effect sizes that can be expected across a range of indicators. Ex ante simulations can also be used to look at initial measures of cost-benefit or cost-effectiveness and to compare alternative interventions for generating changes in the outcomes of interest.

A clearly articulated results chain provides a useful map for selecting the indicators that will be measured along the chain. They will include indicators used both to monitor program implementation and to evaluate results. Again, it is useful to engage program stakeholders from both the policy and research teams in selecting these indicators, to ensure that those selected are good measures of program performance. A widely used rule of thumb to ensure that the indicators used are good measures is summed up by the acronym SMART. Indicators should be the following:

Key Concept

Good indicators are SMART (specific, measurable, attributable, realistic, and targeted).

- *Specific:* To measure the information required as closely as possible
- *Measurable:* To ensure that the information can be readily obtained

- *Attributable:* To ensure that each measure is linked to the project's efforts
- *Realistic:* To ensure that the data can be obtained in a timely fashion, with reasonable frequency, and at reasonable cost
- *Targeted:* To the objective population.

When choosing indicators, remember that it is important to identify indicators all along the results chain, and not just at the level of outcomes, so that you will be able to track the causal logic of any program outcomes that are observed. Indeed, with implementation evaluations that focus on testing two or more design alternatives, the results of interest may occur earlier in the results chain, as an earlier output or early-stage outcome. Even if you are only interested in outcome measures for evaluation, it is still important to track implementation indicators, so you can determine whether interventions have been carried out as planned, whether they have reached their intended beneficiaries, and whether they have arrived on time. Without these indicators all along the results chain, the impact evaluation risks producing a "black box" that identifies whether or not the predicted results materialized; however, it will not be able to explain why that was the case.

Checklist: Getting Data for Your Indicators

As a final checklist once indicators are selected, it is useful to consider the arrangements for producing the data to measure the indicators. A full discussion of where to get data for your evaluation is provided in Section 4. This checklist covers practical arrangements needed to ensure that you can produce each of the indicators reliably and on time (adapted from UNDP 2009):

- ✓ Are the indicators (outputs and outcomes) clearly specified? These are drawn from the core evaluation questions and should be consistent with program design documents and the results chain.
- ✓ Are the indicators SMART (specific, measurable, attributable, realistic, and targeted)?
- ✓ What is the source of data for each of the indicators? There needs to be clarity on the source from which data will be obtained, such as a survey, a review, or administrative data.
- ✓ With what frequency will data be collected? Include a timeline.
- ✓ Who is responsible for collecting the data? Delineate who is responsible for organizing the data collection, verifying data quality and source, and ensuring compliance with ethical standards.

- ✓ Who is responsible for analysis and reporting? Specify the frequency of analysis, analysis method, and responsibility for reporting.
- ✓ What resources are needed to produce the data? Ensure that the resources required are clear and committed to producing the data, which is often the most expensive part of an evaluation if collecting primary data.
- ✓ Is there appropriate documentation? Plans should be in place for how the data will be documented, including using a registry and ensuring anonymity.
- ✓ What are the risks involved? Consider the risks and assumptions in carrying out the planned monitoring and evaluation activities, and how they might affect the timing and quality of the data and of the indicators.

Additional Resources

- For accompanying material to this chapter and hyperlinks to additional resources, please see the Impact Evaluation in Practice website (www.worldbank.org/ieinpractice).
- A theory of change figure, a results chain template, and examples of indicators for results-based financing are presented in Module 1 of the World Bank's Impact Evaluation Toolkit (www.worldbank.org/health/impactevaluationtoolkit).
- A good review of theories of change appears in Imas, Linda G. M., and Ray C. Rist. 2009. *The Road to Results: Designing and Conducting Effective Development Evaluations.* Washington, DC: World Bank.
- For discussions on how to select performance indicators, see the following:
  - Imas, Linda G. M., and Ray C. Rist. 2009. *The Road to Results: Designing and Conducting Effective Development Evaluations.* Washington, DC: World Bank.
  - Kusek, Jody Zall, and Ray C. Rist. 2004. *Ten Steps to a Results-Based Monitoring and Evaluation System.* Washington, DC: World Bank.

References

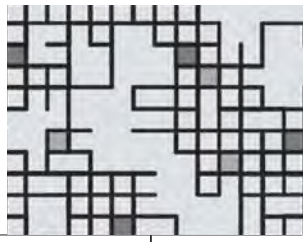
Cattaneo, Matias, Sebastian Galiani, Paul Gertler, Sebastian Martinez, and Rocio Titiunik. 2009. "Housing, Health and Happiness." *American Economic Journal: Economic Policy* 1 (1): 75–105.

Imas, Linda G. M., and Ray C. Rist. 2009. *The Road to Results: Designing and Conducting Effective Development Evaluations.* Washington, DC: World Bank.

Kusek, Jody Zall, and Ray C. Rist. 2004. *Ten Steps to a Results-Based Monitoring and Evaluation System.* Washington, DC: World Bank.

Ludwig, Jens, Jeffrey R. Kling, and Sendhil Mullainathan. 2011. "Mechanism Experiments and Policy Evaluations." *Journal of Economic Perspectives* 25 (3): 17–38.

UNDP (United Nations Development Programme). 2009. *Handbook on Planning, Monitoring and Evaluating for Development Results.* New York: UNDP.



Part 2

HOW TO EVALUATE

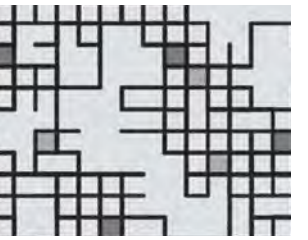
Part 2 of this book explains what impact evaluations do, what questions they answer, what methods are available for conducting them, and the advantages and disadvantages of each. The approach to impact evaluation advocated in this book favors the selection of the most rigorous method compatible with a program's operational characteristics. The menu of impact evaluation options discussed includes randomized assignment, instrumental variables, regression discontinuity design, difference-in-differences, and matching. All of these approaches share the objective of constructing valid comparison groups so that the true impacts of a program can be estimated.

We begin in chapter 3 by introducing the concept of the *counterfactual* as the crux of any impact evaluation, explaining the properties that the estimate of the counterfactual must have, and providing examples of invalid or counterfeit estimates of the counterfactual. Chapters 4–8 then discuss each methodology, covering randomized assignment in chapter 4, instrumental variables in chapter 5, regression discontinuity design in chapter 6, difference-in-differences in chapter 7, and matching in chapter 8. We discuss why and how each method

can produce a valid estimate of the counterfactual, in which policy context each can be implemented, and the main limitations of each method. We illustrate the use of each method using specific real-world examples of impact evaluations that have used these methods, as well as the Health Insurance Subsidy Program (HISP) case study that was introduced in chapter 2. In chapter 9, we discuss how to address problems that can arise during implementation, recognizing that impact evaluations are often not implemented exactly as designed. In this context, we review common challenges including imperfect compliance, spillovers, and attrition, and provide guidance on how to address these issues. Chapter 10 concludes with guidance on evaluations of multifaceted programs, notably those with different treatment levels and multiple treatment arms.

Throughout part 2, you will have a chance to apply methods and test your understanding using the HISP case study. Remember that the key evaluation question for HISP policymakers is, what is the impact of HISP on poor households' out-of-pocket health expenditures? We will use the HISP data set to illustrate each evaluation method and try to answer this question. You should assume that the data have already been properly assembled so as to eliminate any data-related problems. The book will provide you with the results of the analysis, which you will be asked to interpret. Specifically, your task will be to determine why the estimate of the impact of HISP changes with each method and to decide which results are sufficiently reliable to serve as a justification for (or against) expanding HISP. Solutions to the questions are provided on the Impact Evaluation in Practice website (www.worldbank.org/ieinpractice). If you are interested in replicating the analysis, you will also find the data set, analysis code in the Stata software, and a technical companion that provides a more formal treatment of data analysis on this website.

Part 3 begins with how to use the rules of program operation—namely, a program's available resources, criteria for selecting beneficiaries, and timing for implementation—as the basis for selecting an impact evaluation method. A simple framework is set out to determine which of the impact evaluation methodologies presented in part 2 is most suitable for a given program, depending on its operational rules.



CHAPTER 3

Causal Inference and Counterfactuals

Causal Inference

We begin by examining two concepts that are integral to the process of conducting accurate and reliable impact evaluations—causal inference and counterfactuals.

Many policy questions involve cause-and-effect relationships: Does teacher training improve students' test scores? Do conditional cash transfer programs cause better health outcomes in children? Do vocational training programs increase trainees' incomes?

Impact evaluations seek to answer such cause-and-effect questions precisely. Assessing the impact of a program on a set of outcomes is the equivalent of assessing the causal effect of the program on those outcomes.[1]

Although cause-and-effect questions are common, answering them accurately can be challenging. In the context of a vocational training program, for example, simply observing that a trainee's income increases after she has completed such a program is not sufficient to establish causality. The trainee's income might have increased even if she had not taken the training—because of her own efforts, because of changing labor market conditions, or because of many other factors that can affect income. Impact evaluations help us overcome the challenge of establishing causality by empirically establishing to what extent a particular program—*and that program alone*—contributed to

Key Concept

Impact evaluations establish the extent to which a program—and that program alone—caused a change in an outcome.

the change in an outcome. To establish causality between a program and an outcome, we use impact evaluation methods to rule out the possibility that any factors other than the program of interest explain the observed impact.

The answer to the basic impact evaluation question—what is the impact or causal effect of a program (P) on an outcome of interest (Y)?—is given by the basic impact evaluation formula:

$$\Delta = (Y \mid P = 1) - (Y \mid P = 0).$$

This formula states that the causal impact (Δ) of a program (P) on an outcome (Y) is the difference between the outcome (Y) *with* the program (in other words, when $P = 1$) and the same outcome (Y) *without* the program (that is, when $P = 0$).

For example, if P denotes a vocational training program and Y denotes income, then the causal impact of the vocational training program (Δ) is the difference between a person's income (Y) after participation in the vocational training program (in other words, when $P = 1$) and the same person's income (Y) at the same point in time if he or she had not participated in the program (in other words, when $P = 0$). To put it another way, we would like to measure income at the same point in time for the same unit of observation (a person, in this case), but in two different states of the world. If it were possible to do this, we would be observing how much income the same individual would have had at the same point in time both with and without the program, so that the *only* possible explanation for any difference in that person's income would be the program. By comparing the same individual with herself at the same moment, we would have managed to eliminate any outside factors that might also have explained the difference in outcomes. We could then be confident that the relationship between the vocational training program and the change in income is causal.

The basic impact evaluation formula is valid for any unit that is being analyzed—a person, a household, a community, a business, a school, a hospital, or other unit of observation that may receive or be affected by a program. The formula is also valid for any outcome (Y) that is related to the program at hand. Once we measure the two key components of this formula—the outcome (Y) both with the program and without it—we can answer any question about the program's impact.

The Counterfactual

As discussed, we can think of the impact (Δ) of a program as the difference in outcomes (Y) for the same unit (person, household, community, and so on) with and without participation in a program. Yet we know

that measuring the same unit in two different states at the same time is impossible. At any given moment in time, a unit either participated in the program or did not participate. The unit cannot be observed simultaneously in two different states (in other words, with and without the program). This is called the *counterfactual problem*: How do we measure what would have happened if the other circumstance had prevailed? Although we can observe and measure the outcome (Y) for a program participant ($Y \mid P = 1$), there are no data to establish what her outcome would have been in the absence of the program ($Y \mid P = 0$). In the basic impact evaluation formula, *the term* ($Y \mid P = 0$) *represents the counterfactual*. We can think of this as *what would have happened* to the outcome if a person or unit of observation had not participated in the program.

Key Concept

The counterfactual is what would have happened—what the outcome (Y) would have been for a program participant—in the absence of the program (P).

For example, imagine that "Mr. Unfortunate" takes a pill and then dies five days later. Just because Mr. Unfortunate died after taking the pill, you cannot conclude that the pill *caused* his death. Maybe he was very sick when he took the pill, and it was the illness that caused his death, rather than the pill. Inferring causality will require that you rule out other potential factors that could have affected the outcome under consideration. In the simple example of determining whether taking the pill caused Mr. Unfortunate's death, an evaluator would need to establish what would have happened to Mr. Unfortunate if he had *not* taken the pill. Since Mr. Unfortunate did in fact take the pill, it is not possible to observe directly what would have happened if he had not done so. What would have happened to him if he had not taken the pill is the counterfactual. In order to identify the impact of the pill, the evaluator's main challenge is determining what the counterfactual state of the world for Mr. Unfortunate actually looks like (see box 3.1 for another example).

Key Concept

Since we cannot directly observe the counterfactual, we must estimate it.

When conducting an impact evaluation, it is relatively straightforward to obtain the first term of the basic formula ($Y \mid P = 1$)—the outcome with a program (also known as *under treatment*). We simply measure the outcome of interest for the program participant. However, we cannot directly observe the second term of the formula ($Y \mid P = 0$) for the participant. We need to fill in this missing piece of information by *estimating the counterfactual*.

To help us think through this key concept of estimating the counterfactual, we turn to another hypothetical example. Solving the counterfactual problem would be possible if the evaluator could find a "perfect clone" for a program participant (figure 3.1). For example, let us say that Mr. Fulanito starts receiving US$12 in pocket money allowance, and we want to measure the impact of this treatment on his consumption of candies. If you could identify a perfect clone for Mr. Fulanito, the evaluation would be easy: you could

Box 3.1: The Counterfactual Problem: "Miss Unique" and the Cash Transfer Program

"Miss Unique" is a newborn baby girl whose mother is offered a monthly cash transfer so long as she ensures that Miss Unique receives regular health checkups at the local health center, that she is immunized, and that her growth is monitored. The government posits that the cash transfer will motivate Miss Unique's mother to seek the health services required by the program and will help Miss Unique grow strong and tall. For its impact evaluation of the cash transfer, the government selects height as an outcome indicator for long-term health.

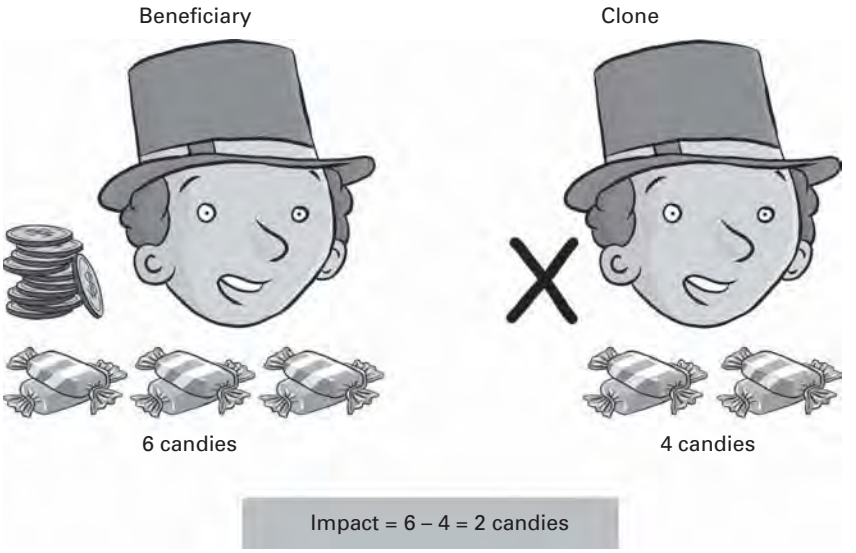
Assume that you are able to measure Miss Unique's height at the age of 3. Ideally, to evaluate the impact of the program, you would want to measure Miss Unique's height at the age of 3 with her mother having received the cash transfer, and also Miss Unique's height at the age of 3 had her mother not received the cash transfer. You would then compare the two heights to establish impact. If you were able to compare Miss Unique's height at the age of 3 with the program to Miss Unique's height at the age of 3 without the program, you would know that any difference in height had been caused only by the cash transfer program. Because everything else about Miss Unique would be the same, there would be no other characteristics that could explain the difference in height.

Unfortunately, however, it is impossible to observe Miss Unique both with and without the cash transfer program: either her family follows the conditions (checkups, immunizations, growth monitoring) and receives the cash transfer or it does not. In other words, we cannot observe what the counterfactual is. Since Miss Unique's mother actually followed the conditions and received the cash transfer, we cannot know how tall Miss Unique would have been had her mother not received the cash transfer.

Finding an appropriate comparison for Miss Unique will be challenging because she is, precisely, unique. Her exact socioeconomic background, genetic attributes, and personal and household characteristics cannot be found in anybody else. If we were simply to compare Miss Unique with a child who is not enrolled in the cash transfer program—say, "Mr. Inimitable"—the comparison may not be adequate. Miss Unique cannot be exactly identical to Mr. Inimitable. Miss Unique and Mr. Inimitable may not look the same, they may not live in the same place, they may not have the same parents, and they may not have been the same height when they were born. So if we observe that Mr. Inimitable is shorter than Miss Unique at the age of 3, we cannot know whether the difference is due to the cash transfer program or to one of the many other differences between these two children.

just compare the number of candies eaten by Mr. Fulanito (say, 6) when he receives the pocket money with the number of candies eaten by his clone (say, 4), who receives no pocket money. In this case, the impact of the pocket money would be 2 candies: the difference between the number of candies consumed under treatment (6) and the number of candies consumed

Figure 3.1 The Perfect Clone



without treatment (4). In reality, we know that it is impossible to identify perfect clones: even between genetically identical twins, there are important differences.

Estimating the Counterfactual

The key to estimating the counterfactual for program participants is to move from the individual or unit level to the group level. Although no perfect clone exists for a single unit, we can rely on statistical properties to generate two *groups* of units that, if their numbers are large enough, are statistically indistinguishable from each other at the group level. The group that participates in the program is known as the *treatment group,* and its outcome is $(Y \mid P = 1)$ after it has participated in the program. The statistically identical *comparison group* (sometimes called the *control group*) is the group that remains unaffected by the program, and allows us to estimate the counterfactual outcome $(Y \mid P = 0)$: that is, the outcome that would have prevailed for the treatment group had it not received the program.

So in practice, the challenge of an impact evaluation is to identify a treatment group and a comparison group that are statistically identical, on average, in the absence of the program. If the two groups are identical, with the sole exception that one group participates in the program

Key Concept

Without a comparison group that yields an accurate estimate of the counterfactual, the true impact of a program cannot be established.

and the other does not, then we can be sure that any difference in outcomes must be due to the program. Finding such comparison groups is the crux of any impact evaluation, regardless of what type of program is being evaluated. Simply put, without a comparison group that yields an accurate estimate of the counterfactual, the true impact of a program cannot be established.

The main challenge for identifying impacts, then, is to find *a valid comparison group* that has the same characteristics as the treatment group in the absence of a program. Specifically, the treatment and comparison groups must be the same in at least three ways.

First, the average characteristics of the treatment group and the comparison group must be identical in the absence of the program.[2] Although it is not necessary that individual units in the treatment group have "perfect clones" in the comparison group, *on average* the characteristics of treatment and comparison groups should be the same. For example, the average age of units in the treatment group should be the same as in the comparison group.

Second, the treatment should not affect the comparison group either directly or indirectly. In the pocket money example, the treatment group should not transfer resources to the comparison group (direct effect) or affect the price of candy in the local markets (indirect effect). For example, if we want to isolate the impact of pocket money on candy consumption, the treatment group should not also be offered more trips to the candy store than the comparison group; otherwise, we would be unable to distinguish whether additional candy consumption is due to the pocket money or to the extra trips to the store.

Key Concept

A valid comparison group (1) has the same characteristics, on average, as the treatment group in the absence of the program; (2) remains unaffected by the program; and (3) would react to the program in the same way as the treatment group, if given the program.

Third, the outcomes of units in the control group should change the same way as outcomes in the treatment group, if both groups were given the program (or not). In this sense, the treatment and comparison groups should react to the program in the same way. For example, if incomes of people in the treatment group increased by US$100 thanks to a training program, then incomes of people in the comparison group would have also increased by US$100, had they been given training.

When these three conditions are met, then only the existence of the program of interest will explain any differences in the outcome (Y) between the two groups. This is because the only difference between the treatment and comparison groups is that the members of the treatment group receive the program, while the members of the comparison group do not. When the difference in outcome can be entirely attributed to the program, the causal impact of the program has been identified.

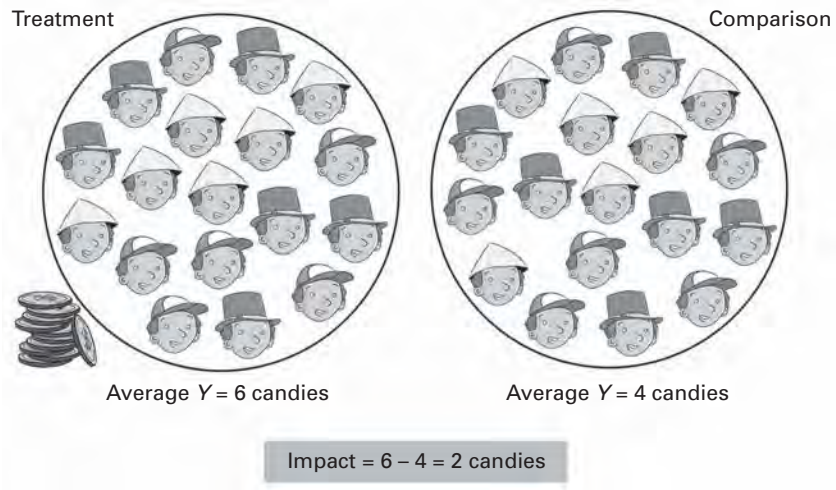
Returning to the case of Mr. Fulanito, we saw that in order to estimate the impact of pocket money on his consumption of candies would require the implausible task of finding Mr. Fulanito's perfect clone. Instead of looking at the impact solely for one individual, it is more realistic to look at the average impact for a group of individuals (figure 3.2). If you could identify another group of individuals that shares the same average age, gender composition, education, preference for candy, and so on, except that it does not receive additional pocket money, then you could estimate the pocket money's impact. This would simply be the difference between the average consumption of candies in the two groups. Thus if the *treatment group* consumes an average of 6 candies per person, while the *comparison group* consumes an average of 4, the average impact of the additional pocket money on candy consumption would be 2 candies.

Having defined a *valid comparison group*, it is important to consider what would happen if we decided to go ahead with an evaluation without finding such a group. Intuitively, an invalid comparison group is one that differs from the treatment group in some way other than the absence of the treatment. Those additional differences can cause the estimate of impact to be invalid or, in statistical terms, *biased*: the impact evaluation will not estimate the true impact of the program. Rather, it will estimate the effect of the program mixed with those other differences.

Key Concept

When the comparison group does not accurately estimate the true counterfactual, then the estimated impact of the program will be invalid. In statistical terms, it will be *biased*.

Figure 3.2 A Valid Comparison Group



Two Counterfeit Estimates of the Counterfactual

In the remainder of part 2 of this book, we will discuss the various methods that can be used to construct valid comparison groups that will allow you to estimate the counterfactual. Before doing so, however, it is useful to discuss two common, but highly risky, methods of constructing comparison groups that many times lead to inappropriate ("counterfeit") estimates of the counterfactual:

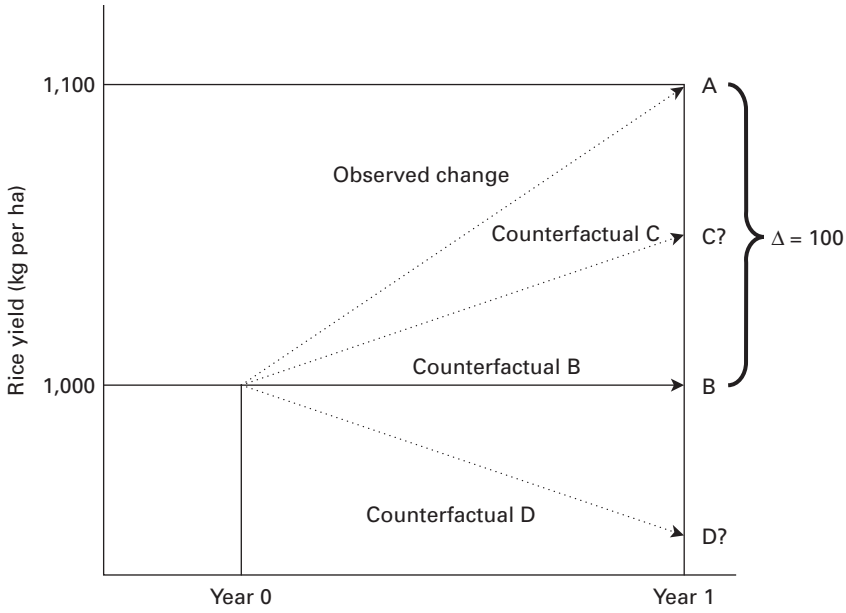
- *Before-and-after comparisons* (also known as *pre-post* or *reflexive comparisons*) compare the outcomes of the same group before and after participating in a program.
- *Enrolled-and-nonenrolled* (or *self-selected*) *comparisons* compare the outcomes of a group that chooses to participate in a program with those of a group that chooses not to participate.

Counterfeit Counterfactual Estimate 1: Comparing Outcomes Before and After a Program

A before-and-after comparison attempts to establish the impact of a program by tracking changes in outcomes for program participants over time. Returning to the basic impact evaluation formula, the outcome for the treatment group ($Y \mid P = 1$) is simply the outcome after participating in the program. However, before-and-after comparisons take the estimated counterfactual ($Y \mid P = 0$) as the outcome for the treatment group *before* the intervention started. In essence, this comparison assumes that if the program had never existed, the outcome (Y) for program participants would have been exactly the same as their situation before the program. Unfortunately, for a majority of programs implemented over a series of months or years, this assumption simply does not hold.

Consider the evaluation of a microfinance program for poor, rural farmers. The program provides microloans to farmers to enable them to buy fertilizer to increase their rice production. You observe that in the year before the program starts, farmers harvested an average of 1,000 kilograms (kg) of rice per hectare (point B in figure 3.3). The microfinance scheme is launched, and a year later rice yields have increased to 1,100 kg per hectare (point A in figure 3.3). If you were trying to evaluate impact using a before-and-after comparison, you would use the baseline outcome as an estimate of the counterfactual. Applying the basic impact evaluation formula, you would conclude that the program had increased rice yields by 100 kg per hectare ($A-B$).

Figure 3.3 Before-and-After Estimates of a Microfinance Program



Note: Δ = Change in rice yield (kg); ha = hectares; kg = kilograms.

However, imagine that rainfall was normal in the year before the program was launched, but a drought occurred in the year the program operated. Because of the drought, the farmers' average yield without the microloan scheme is likely to be lower than *B:* say, at level *D*. In that case, the true impact of the program would be *A*–*D*, which is larger than the 100 kg estimated using the before-and-after comparison. By contrast, if rainfall actually improved between the two years, the counterfactual rice yield might have been at level *C*. In that case, the true program impact would have been smaller than 100 kg. In other words, unless our impact analysis can account for rainfall and *every other factor* that can affect rice yields over time, we simply cannot calculate the true impact of the program by making a before-and-after comparison.

In the previous microfinance example, rainfall was one of myriad outside factors which might affect the program's outcome of interest (rice yields) over time. Likewise, many of the outcomes that development programs aim to improve, such as income, productivity, health, or education, are affected by an array of factors over time. For that reason, the baseline outcome is almost never a good estimate of the counterfactual. That is why we consider it a counterfeit estimate of the counterfactual.

Evaluating the Impact of HISP: Doing a Before-and-After Comparison of Outcomes

Recall that the Health Insurance Subsidy Program (HISP) is a new program in your country that subsidizes the purchase of health insurance for poor rural households and that this insurance covers expenses related to health care and medicine for those enrolled. The objective of HISP is to reduce what poor households spend on primary care and medicine and ultimately to improve health outcomes. Although many outcome indicators could be considered for the program evaluation, your government is particularly interested in analyzing the effects of HISP on per capita yearly out-of-pocket expenditures (subsequently referred to simply as *health expenditures*).

HISP will represent a hefty proportion of the national budget if scaled up nationally—up to 1.5 percent of gross domestic product (GDP) by some estimates. Furthermore, substantial administrative and logistical complexities are involved in running a program of this nature. For these reasons, a decision has been made at the highest levels of government to introduce HISP first as a pilot program and then, depending on the results of the first phase, to scale it up gradually over time. Based on the results of financial and cost-benefit analyses, the president and her cabinet have announced that for HISP to be viable and to be extended nationally, it must reduce yearly per capita health expenditures of poor rural households by at least US$10 on average, compared to what they would have spent in the absence of the program, and it must do so within two years.

HISP will be introduced in 100 rural villages during the initial pilot phase. Just before the start of the program, your government hires a survey firm to conduct a baseline survey of all 4,959 households in these villages. The survey collects detailed information on every household, including their demographic composition, assets, access to health services, and health expenditures in the past year. Shortly after the baseline survey is conducted, HISP is introduced in the 100 pilot villages with great fanfare, including community events and other promotional campaigns to encourage households to enroll.

Of the 4,959 households in the baseline sample, a total of 2,907 enroll in HISP, and the program operates successfully over the next two years. All health clinics and pharmacies serving the 100 villages accept patients with the insurance scheme, and surveys show that most enrolled households are satisfied with the program. At the end of the two-year pilot period, a second round of evaluation data is collected on the same sample of 4,959 households.[3]

The president and the minister of health have put you in charge of overseeing the impact evaluation for HISP and recommending whether or not to extend the program nationally. Your impact evaluation question of interest is, what is the impact of HISP on poor households' out-of-pocket health expenditures? Remember that the stakes are high. If HISP is found to reduce health expenditures by US$10 or more, it will be extended nationally. If the program did not reach the US$10 target, you will recommend against scaling it up.

The first "expert" consultant you hire indicates that to estimate the impact of HISP, you must calculate the change in health expenditures over time for the households that enrolled. The consultant argues that because HISP covers all health costs, any decrease in expenditures over time must be attributable to the effect of HISP. Using the subset of enrolled households, you calculate their average health expenditures before the implementation of the program and then again two years later. In other words, you perform a before-and-after comparison. The results are shown in table 3.1. You observe that the treatment group reduced its out-of-pocket health expenditures by US$6.65, from US$14.49 before the introduction of HISP to US$7.84 two years later. As denoted by the value of the t-statistic (*t-stat*), the difference between health expenditures before and after the program is *statistically significant*.[4] This means that you find strong evidence against the claim that the true difference between expenditures before and after the intervention is zero.

Even though the before-and-after comparison is for the same group of households, you are concerned that other circumstances may have also changed for these households over the past two years, affecting their health expenditures. For example, a number of new drugs have recently become available. You are also concerned that the reduction in health expenditures may have resulted in part from the financial crisis that your country recently experienced. To address some of these concerns, your consultant conducts a more sophisticated *regression analysis* that will try to control for some additional factors.

Table 3.1 Evaluating HISP: Before-and-After Comparison

	After	Before	Difference	*t*-stat
Household health expenditures (US$)	7.84	14.49	–6.65**	–39.76

Note: Significance level: ** = 1 percent.

Table 3.2 Evaluating HISP: Before-and-After with Regression Analysis

	Linear regression	Multivariate linear regression
Estimated impact on household health expenditures (US$)	–6.65** (0.23)	–6.71** (0.23)

Note: Standard errors are in parentheses. Significance level: ** = 1 percent.

Regression analysis uses statistics to analyze the relationships between a dependent variable (the variable to be explained) and explanatory variables. The results appear in table 3.2. A linear regression is the simplest form: the dependent variable is health expenditures, and there is only one explanatory variable: a binary (0–1) indicator that takes the value 0 if the observation is taken at baseline and 1 if the observation is taken at follow-up.

A multivariate linear regression adds explanatory variables to *control for*, or *hold constant*, other characteristics that are observed for the households in your sample, including indicators for wealth (assets), household composition, and so on.[5]

You note that the result from the linear regression is equivalent to the simple before-and-after difference in average health expenditures from table 3.1 (a reduction of US$6.65 in health expenditures). Once you use multivariate linear regression to control for other factors available in your data, you find a similar result—a decrease of US$6.71 in health expenditures.

HISP Question 1

A. Does the before-and-after comparison control for all the factors that affect health expenditures over time?

B. Based on these results produced by the before-and-after analysis, should HISP be scaled up nationally?

Counterfeit Counterfactual Estimate 2: Comparing Enrolled and Nonenrolled (Self-Selected) Groups

Comparing a group of individuals that voluntarily signs up for a program to a group of individuals that *chooses* not participate is another risky approach to evaluating impact. A comparison group that *self-selects* out of a program will provide another counterfeit counterfactual estimate. *Selection* occurs when program participation is based on the preferences, decisions, or

unobserved characteristics of potential participants.

Consider, for example, a vocational training program for unemployed youth. Assume that two years after the program has been launched, an evaluation attempts to estimate its impact on income by comparing the average incomes of a group of youth who chose to enroll in the program versus a group of youth who, despite being eligible, chose not to enroll. Assume that the results show that youth who chose to enroll in the program make twice as much as those who chose not to enroll. How should these results be interpreted? In this case, the counterfactual is estimated based on the incomes of individuals who decided not to enroll in the program. Yet the two groups are likely to be fundamentally different. Those individuals who chose to participate may be highly motivated to improve their livelihoods and may expect a high return to training. In contrast, those who chose not to enroll may be discouraged youth who do not expect to benefit from this type of program. It is likely that these two types would perform quite differently in the labor market and would have different incomes even without the vocational training program.

The same issue arises when admission to a program is based on unobserved preferences of program administrators. Say, for example, that the program administrators base admission and enrollment on an interview. Those individuals who are admitted to the program might be those who the administrators think have a good chance of benefiting from the program. Those who are not admitted might show less motivation at the interview, have lower qualifications, or just lack good interview skills. Again, it is likely that these two groups of young people would have different incomes in the labor market even in absence of a vocational training program.

Thus the group that did not enroll does not provide a good estimate of the counterfactual. If you observe a difference in incomes between the two groups, you will not be able to determine whether it comes from the training program or from the underlying differences in motivation, skills, and other factors that exist between the two groups. The fact that less motivated or less qualified individuals did not enroll in the training program therefore leads to a bias in the program's impact.[6] This bias is called *selection bias*. More generally, selection bias will occur when the reasons for which an individual participates in a program are correlated with outcomes, even in absence of the program. Ensuring that the estimated impact is free of selection bias is one of the major objectives and challenges for any impact evaluation. In this example, if the young people who enrolled in vocational training would have had higher incomes even in the absence of the program, the selection bias would be positive; in other words, you would overestimate the impact of the vocational training program by attributing to the program the higher incomes that participants would have had anyway.

Key Concept

Selection bias occurs when the reasons for which an individual participates in a program are correlated with outcomes. Ensuring that the estimated impact is free of selection bias is one of the major objectives and challenges for any impact evaluation.

Evaluating the Impact of HISP: Comparing Enrolled and Nonenrolled Households

Having thought through the before-and-after comparison a bit further with your evaluation team, you realize that there are still many other factors that can explain part of the change in health expenditures over time (in particular, the minister of finance is concerned that a recent financial crisis may have affected households' income, and may explain the observed change in health expenditures).

Another consultant suggests that it would be more appropriate to estimate the counterfactual in the post-intervention period: that is, two years after the program started. The consultant correctly notes that of the 4,959 households in the baseline sample, only 2,907 actually enrolled in the program, so approximately 41 percent of the households in the sample remain without HISP coverage. The consultant argues that all households within the 100 pilot villages were eligible to enroll. These households all share the same health clinics and are subject to the same local prices for pharmaceuticals. Moreover, most households are engaged in similar economic activities. The consultant argues that in these circumstances, the outcomes of the nonenrolled group after the intervention could serve to estimate the counterfactual outcome of the group enrolled in HISP. You therefore decide to calculate average health expenditures in the post-intervention period for both the households that enrolled in the program and the households that did not. The results are shown in table 3.3. Using the average health expenditures of the nonenrolled households as the estimate of the counterfactual, you find that the program has reduced average health expenditures by approximately US$14.46.

When discussing this result further with the consultant, you raise the question of whether the households that chose not to enroll in the program may be systematically different from the ones that did enroll. For example, the households that signed up for HISP may be ones that

Table 3.3 Evaluating HISP: Enrolled-Nonenrolled Comparison of Means

	Enrolled	Nonenrolled	Difference	*t*-stat
Household health expenditures (US$)	7.84	22.30	−14.46**	−49.08

Note: Significance level: ** = 1 percent.

Table 3.4 Evaluating HISP: Enrolled-Nonenrolled Regression Analysis

	Linear regression	Multivariate linear regression
Estimated impact on household health expenditures (US$)	−14.46** (0.33)	−9.98** (0.29)

Note: Standard errors are in parentheses. Significance level: ** = 1 percent.

expected to have higher health expenditures, or people who were better informed about the program, or people who care more for the health of their families. Alternatively, perhaps the households that enrolled were poorer, on average, than those who did not enroll, given that HISP was targeted to poor households. Your consultant argues that regression analysis can control for these potential differences between the two groups. She therefore carries out an additional multivariate regression that controls for all the household characteristics that she can find in the data set, and estimates the impact of the program as shown in table 3.4.

With a simple linear regression of health expenditures on an indicator variable of whether or not a household enrolled in the program, you find an estimated impact of minus US$14.46; in other words, you estimate that the program has decreased average health expenditures by US$14.46. However, when all other characteristics in the data are controlled for, you estimate that the program has reduced health expenditures by US$9.98 per year.

HISP Question 2

A. Does this analysis likely control for all the factors that determine differences in health expenditures between the two groups?

B. Based on these results produced by the enrolled-nonenrolled method, should HISP be scaled up nationally?

Additional Resources

- For accompanying material to the book and hyperlinks to additional resources, please see the Impact Evaluation in Practice website (www.worldbank.org/ieinpractice).

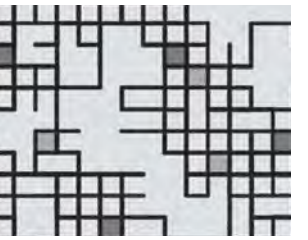
Notes

1. We use the Rubin Causal Model as a framework for causal inference (Imbens and Rubin 2008; Rubin 1974).
2. This condition will be relaxed in some impact evaluation methods, which will require instead that the average *change* in outcomes (trends) is the same in the absence of the program.
3. We are assuming that no households have left the sample over two years (there is zero sample attrition). This is not a realistic assumption for most household surveys. In practice, families that move sometimes cannot be tracked to their new location, and some households break up and cease to exist altogether.
4. Note that a *t*-statistic (*t*-stat) of 1.96 or more (in absolute value) is statistically significant at the 5 percent level.
5. For more on multivariate analysis, see the online technical companion on the Impact Evaluation in Practice website (www.worldbank.org/ieinpractice).
6. Another example, if youth who anticipate benefiting considerably from the training scheme are also more likely to enroll (for example, because they anticipate higher wages with training), then comparing them to a group with lower expected returns that does not enroll will yield a biased estimate of impact.

References

Imbens, Guido W., and Donald B. Rubin. 2008. "Rubin Causal Model." In *The New Palgrave Dictionary of Economics*, second edition, edited by Steven N. Durlauf and Lawrence E. Blume. Palgrave.

Rubin, Donald B. 1974. "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies." *Journal of Educational Psychology* 66 (5): 688–701.



CHAPTER 4

Randomized Assignment

Evaluating Programs Based on the Rules of Assignment

Having discussed two "counterfeit" estimates of the counterfactual that are commonly used but have a high risk of bias—before-and-after comparisons and enrolled-nonenrolled comparisons—we now turn to a set of methods that can be applied to estimate program impacts more accurately. Such estimation, however, is not always as straightforward as it might seem at first glance. Most programs are designed and implemented in a complex and changing environment in which many factors can influence outcomes for both program participants and those who do not participate. Droughts, earthquakes, recessions, changes in government, and changes in international and local policies are all part of the real world. In an evaluation, we want to make sure that the estimated impact of our program remains valid despite these myriad factors.

A program's rules for selecting participants will be the key parameter for determining the impact evaluation method. We believe that in most cases, the evaluation methods should try to fit within the context of a program's operational rules (with a few tweaks here and there)—and not the other way around. However, we also start from the premise that *all programs should have fair and transparent rules for program assignment*. One of the fairest and most transparent rules for allocating scarce resources among equally

deserving populations turns out to be giving everyone who is eligible an equal opportunity to participate in the program. One way to do that is simply to run a lottery.

In this chapter, we will examine a method that is akin to running a lottery that decides who enters a program at a given time and who does not: the *randomized assignment method,* also known as *randomized controlled trials* (RCTs). This method not only provides program administrators with a fair and transparent rule for allocating scarce resources *among equally deserving populations,* but also represents the strongest method for evaluating the impact of a program. Thus the application of this method to evaluate impacts of social programs has increased substantially in recent years.

Randomized Assignment of Treatment

When a program is assigned at random—that is, using a lottery—over a large eligible population, we can generate a robust estimate of the counterfactual. *Randomized assignment* of treatment is considered the gold standard of impact evaluation. It uses a random process, or chance, to decide who is granted access to the program and who is not.[1] Under randomized assignment, every eligible unit (for example, an individual, household, business, school, hospital, or community) has the same probability of being selected for treatment by a program.[2]

Before we discuss how to implement randomized assignment in practice and why it generates a strong estimate of the counterfactual, let us take a few moments to consider why randomized assignment is also a fair and transparent way to assign scarce program resources. Once a target population has been defined (say, households below the poverty line, children under the age of 5, or roads in rural areas in the north of the country), randomized assignment is a fair allocation rule because it allows program managers to ensure that every eligible unit has the same chance of receiving the program and that the program is not being assigned using arbitrary or subjective criteria, or even through patronage or other unfair practices. When excess demand for a program exists, randomized assignment is a rule that can be easily explained by program managers, is understood by key constituents, and is considered fair in many circumstances. In addition, when the assignment process is conducted openly and transparently, it cannot easily be manipulated, and therefore it shields program managers from potential accusations of favoritism or corruption. Randomized assignment thus has its own merits as a rationing mechanism that go well beyond its utility as an impact evaluation tool.

In fact, a number of programs routinely use lotteries as a way to select participants from the pool of eligible individuals, primarily because of their advantages for administration and governance.[3] Box 4.1 presents two such cases from Africa.

Randomized assignment can often be derived from a program's operational rules. For many programs, the population of intended

Box 4.1: Randomized Assignment as a Valuable Operational Tool

Randomized assignment can be a useful rule for assigning program benefits, even outside the context of an impact evaluation. The following two cases from Africa illustrate how.

In Côte d'Ivoire, following a period of crisis, the government introduced a temporary employment program that was initially targeted at former combatants and later expanded to youth more generally. The program provided youth with short-term employment opportunities, mostly to clean or rehabilitate roads through the national roads agency. Youth in participating municipalities were invited to register. Given the attractiveness of the benefits, many more youth applied than places where available. In order to come up with a transparent and fair way of allocating the benefits among applicants, program implementers put in place a public lottery process. Once registration had closed and the number of applicants (say *N*) in a location was known, a public lottery was organized. All applications were called to a public location, and small pieces of paper with numbers from 1 to *N* were put in a box. Applicants would then be called one by one to come and draw a number from the box in front of all other applications. Once the number was drawn, it would be read aloud. After all applicants were called, someone would check the remaining numbers in the box one by one to ensure that they were applicants who did not come to the lottery. If *N* spots were available for the program, the applicants having drawn the lowest numbers were selected for the program. The lottery process was organized separately for men and women. The public lottery process was well accepted by participants, and helped provide an image of fairness and transparency to the program in a post-conflict environment marked by social tensions. After several years of operations, researchers used this allocation rule, already integrated in the program operation, to undertake its impact evaluation.

In Niger, the government started to roll out a national safety net project in 2011 with support from the World Bank. Niger is one of the poorest countries in the world, and the population of poor households deserving the program greatly exceeded the available benefits during the first years of operation. Program implementers relied on geographical targeting to identify the departments and communes where the cash transfer program would be implemented first. This could be done, as data existed to determine the relative poverty or vulnerability status of the various departments or communes. However, within communes, very limited

(continued)

Box 4.1: Randomized Assignment as a Valuable Operational Tool *(continued)*

data were available to assess which villages would be more deserving than others based on objective criteria. For the first phase of the project, program implementers decided to use public lotteries to select beneficiary villages within targeted communes. This decision was made in part because the available data to prioritize villages objectively were limited, and in part because an impact evaluation was being embedded in the project. For the public lotteries, all the village chiefs were invited in the municipal center, and the names of their villages were written on a piece of paper, and put in a box. A child would then randomly draw beneficiary villages from the box until the quotas were filled. The procedure was undertaken separately for sedentary and nomadic villages to ensure representation of each group. (After villages were selected, a separate household-level targeting mechanism was implemented to identify the poorest households, which were later enrolled as beneficiaries.) The transparency and fairness of the public lottery was greatly appreciated by local and village authorities, as well as by program implementers—so much that the public lottery process continued to be used in the second and third cycle of the project to select over 1,000 villages throughout the country. Even though the public lottery was not necessary for an impact evaluation at that point, its value as a transparent, fair, and widely accepted operational tool to allocate benefits among equally deserving populations justified its continued use in the eyes of program implementers and local authorities.

Sources: Bertrand and others 2016; Premand, Barry, and Smitz 2016.

participants—that is, the set of all units that the program would like to serve—is larger than the number of participants that the program can actually accommodate at a given time. For example, in a single year an education program might be able to provide school materials to 500 schools out of thousands of eligible schools in the country. Or a rural road improvement program may have a goal of paving 250 rural roads, although there are hundreds more that the program would like to improve. Or a youth employment program may have a goal of reaching 2,000 unemployed youth within its first year of operation, although there are tens of thousands of unemployed young people that the program would ultimately like to serve. For a variety of reasons, programs may be unable to reach the entire population of interest. Budgetary constraints may simply prevent administrators from offering the program to all eligible units from the beginning. Even if budgets are available to cover an unlimited number of participants, capacity constraints will sometimes prevent a program from being rolled out to everyone at the same time. For example, in the case of the youth employment training program, the number of unemployed

youth who want vocational training may be greater than the number of slots available in technical colleges during the first year of the program, and that may limit the number who can enroll.

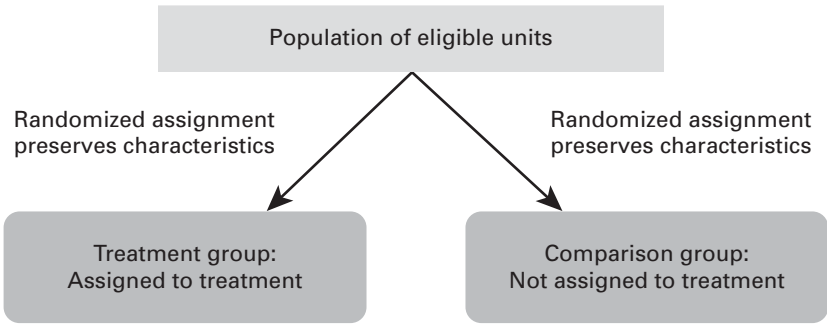
When the population of eligible participants is larger than the number of program places available, someone must make a decision about who will enter the program and who will not. In other words, program administrators must define a rationing mechanism to allocate the program's services. The program could be assigned on a first-come, first-served basis, or based on observed characteristics (for example, serving the poorest areas first); or selection could be based on unobserved characteristics (for example, letting individuals sign up based on their own motivation and knowledge) or on a lottery. Even in contexts where it is possible to rank potential participants based on a measure of need, it may be desirable to allocate some of the benefits by lottery. Take for example a program that targets the poorest 20 percent of households based on a measure of income. If income can only be measured imperfectly, the program could use this measure to include all potential participants who are identified as *extreme poor* (for example the bottom 15 percent). But since income is measured imperfectly, households just below the eligibility threshold at the 20th percentile may or may not be eligible in reality (if we could measure true income), while households just above the 20th percentile may also be eligible or not. In this context, running a lottery to allocate benefits for households around the 20th percentile (for example between the 15th and 25th percentile of the income distribution) could be a fair way to allocate benefits in this group of households.

Why Does Randomized Assignment Produce an Excellent Estimate of the Counterfactual?

As discussed, the ideal comparison group would be as similar as possible to the treatment group in all respects, except with respect to its participation in the program that is being evaluated. When we randomly assign units to treatment and comparison groups, that randomized assignment process in itself will produce two groups that have a high probability of being statistically identical—as long as the number of potential units to which we apply the randomized assignment process is sufficiently large. Specifically, with a large enough number of units, the randomized assignment process will produce groups that have *statistically equivalent averages for all their characteristics*.[4]

Figure 4.1 illustrates why randomized assignment produces a comparison group that is statistically equivalent to the treatment group. Suppose the population of eligible units (the pool of potential participants, or population

Figure 4.1 Characteristics of Groups under Randomized Assignment of Treatment



of interest for the evaluation) consists of 1,000 people. Half are randomly assigned to the treatment group, and the other half are randomly assigned to the comparison group. For example, you could imagine writing the names of all 1,000 people on individual pieces of paper, mixing them up in a bowl, and then asking someone to blindly draw out 500 names. If the first 500 names make up the treatment group, then you would have a randomly assigned treatment group (the first 500 names drawn), and a randomly assigned comparison group (the 500 names left in the bowl).

Now assume that of the original 1,000 people, 40 percent were women. Because the names were selected at random, of the 500 names drawn from the bowl, approximately 40 percent will also be women. If among the 1,000 people, 20 percent had blue eyes, then approximately 20 percent of both the treatment and the comparison groups should have blue eyes, too. In general, if the population of eligible units is large enough, then the randomized assignment mechanism will ensure that any characteristic of the population will transfer to both the treatment group and the comparison group. Just as observed characteristics such as sex or the color of a person's eyes transfer to both the treatment group and the comparison group, then logically characteristics that are more difficult to observe (*unobserved variables*), such as motivation, preferences, or other personality traits that are difficult to measure, would also apply equally to both the treatment and comparison groups. Thus, treatment and comparison groups that are generated through randomized assignment will be similar not only in their observed characteristics but also in their unobserved characteristics. Having two groups that are similar in every way guarantees that the estimated counterfactual approximates the true value of the outcome in the absence of treatment, and that once the program is implemented, the estimated impacts will not suffer from selection bias.

Key Concept

In randomized assignment, each eligible unit has the same probability of being selected for treatment, ensuring equivalence between the treatment and comparison groups in both observed and unobserved characteristics.

When an evaluation uses randomized assignment to treatment and comparison groups, in theory the process should produce two groups that are equivalent, provided it relies on a large enough number of units. With the baseline data from our evaluation sample, we can test this assumption empirically and verify that in fact there are no systematic differences in observed characteristics between the treatment and comparison groups before the program starts. Then, after we launch the program, if we observe differences in outcomes between the treatment and comparison groups, we will know that those differences can be explained only by the introduction of the program, since by construction the two groups were identical at the baseline, before the program started, and are exposed to the same external environmental factors over time. In this sense, the comparison group *controls* for all factors that might also explain the outcome of interest.

To estimate the impact of a program under randomized assignment, we simply take the difference between the outcome under treatment (the mean outcome of the randomly assigned treatment group) and our estimate of the counterfactual (the mean outcome of the randomly assigned comparison group). We can be confident that our estimated impact constitutes the true impact of the program, since we have eliminated all observed and unobserved factors that might otherwise plausibly explain the difference in outcomes. In boxes 4.2 through 4.6, we discuss real world applications of randomized assignment to evaluate the impact of a number of different interventions around the world.

In figure 4.1 we assumed that all units in the eligible population would be assigned to either the treatment group or the comparison group. In some cases, however, it is not necessary to include all units in the evaluation. For example, if the population of eligible units includes 1 million mothers and you want to evaluate the effectiveness of cash bonuses on the probability that they will get their children vaccinated, it may be sufficient to select a representative random sample of, say, 1,000 mothers, and assign those 1,000 to either the treatment group or the comparison group. Figure 4.2 illustrates this process. By the same logic explained above, selecting a random sample from the population of eligible units to form the evaluation sample preserves the characteristics of the population of eligible units. Within the evaluation sample, randomized assignment of individuals to the treatment and comparison groups again preserves the characteristics. We discuss sampling further in chapter 15.

External and Internal Validity

The steps outlined above for randomized assignment of treatment will ensure both the internal and the external validity of the impact estimates (figure 4.2).

Box 4.2: Randomized Assignment as a Program Allocation Rule: Conditional Cash Transfers and Education in Mexico

The Progresa program, now called "Prospera," provides cash transfers to poor mothers in rural Mexico conditional on their children's enrollment in school and regular health checkups (see box 1.1 in chapter 1). The cash transfers, for children in grades 3 through 9, amount to about 50 percent to 75 percent of the private cost of schooling and are guaranteed for three years. The communities and households eligible for the program were determined based on a poverty index created from census data and baseline data collection. Because of a need to phase in the large-scale social program, about two-thirds of the localities (314 out of 495) were randomly selected to receive the program in the first two years, and the remaining 181 served as a comparison group before entering the program in the third year.

Based on the randomized assignment, Schultz (2004) found an average increase in enrollment of 3.4 percent for all students in grades 1–8, with the largest increase among girls who had completed grade 6, at 14.8 percent.[a] The likely reason is that girls tend to drop out of school at greater rates as they get older, so they were given a slightly larger transfer to stay in school past the primary grade levels. These short-term impacts were then extrapolated to predict the longer-term impact of the Progresa program on lifetime schooling and earnings.

Source: Schultz 2004.
a. To be precise, Schultz combined randomized assignment with the difference-in-differences method discussed in chapter 7.

Box 4.3: Randomized Assignment of Grants to Improve Employment Prospects for Youth in Northern Uganda

In 2005, the government of Uganda began a program aimed at decreasing youth unemployment and promoting social stability in the conflict-affected northern region. The Youth Opportunities Program invited groups of young adults to submit grant proposals for business activities and vocational training. Thousands of proposals were submitted, but the government was able to fund only a few hundred.

Taking advantage of the high demand for the program, evaluators worked with the government to randomize which groups received funding. The central government asked district governments to submit more than twice the number of proposals that they could fund. After screening the proposals, the government determined a list of 535 proposals eligible for the program. The proposals were then randomly assigned to the treatment or the comparison group, with 265 in the treatment and 270 in the comparison group.

(continued)

Box 4.3 Randomized Assignment of Grants to Improve Employment Prospects for Youth in Northern Uganda *(continued)*

The grant amount in the treatment group averaged US$382 per person. Four years after the disbursements, youth in the treatment group were more than twice as likely to practice a skilled trade as youth in the comparison group. They also earned 38 percent more and had 57 percent more capital stock. However, researchers found no impact on social cohesion or antisocial behavior.

Source: Blattman, Fiala, and Martinez 2014.

Box 4.4: Randomized Assignment of Water and Sanitation Interventions in Rural Bolivia

Starting in 2012, the Bolivian government, with support from the Inter-American Development Bank, implemented a randomized assignment of water and sanitation interventions for small rural communities. Within the 24 municipalities in the country with the greatest need, the program identified over 369 communities that were eligible for the intervention. Since resources were available to cover only 182 communities, the program used randomized assignment to give each eligible community an equal chance to participate. Together with municipal governments, program administrators organized a series of events where they held public lotteries in the presence of community leaders, the press, and civil society.

First, communities were divided up according to their population size. Then, within each group, community names were drawn at random and placed on a list. The communities that were on the top of the list were assigned to the treatment group. Each lottery was monitored by an independent notary public, who subsequently registered and certified the results, granting an additional level of legitimacy to the process. For communities left out of the program, municipal governments committed to using the same randomly ordered list to allocate future funding after completing the evaluation. In this way, no communities would be left out of the intervention for the sole purposes of the evaluation, but a comparison group would exist so long as budget constraints limited the number of projects in each municipality.

Source: Inter-American Development Bank Project No. BO-L1065, http://www.iadb.org/en/projects/project-description-title,1303.html?id=BO-L1065.
Note: See the public lottery for randomized assignment at https://vimeo.com/86744573.

Internal validity means that the estimated impact of the program is net of all other potential confounding factors—or, in other words, that the comparison group provides an accurate estimate of the counterfactual, so that we are estimating the true impact of the program. Remember that randomized assignment produces a comparison group

Box 4.5: Randomized Assignment of Spring Water Protection to Improve Health in Kenya

The link between water quality and health impacts in developing countries has been well documented. However, the health value of improving infrastructure around water sources is less evident. Kremer and others (2011) measured the effects of a program providing spring protection technology to improve water quality in Kenya, randomly assigning springs to receive the treatment.

Approximately 43 percent of households in rural Western Kenya obtain drinking water from naturally occurring springs. Spring protection technology seals off the source of a water spring to lessen contamination. Starting in 2005, the NGO International Child Support (ICS) implemented a spring protection program in two districts in western Kenya. Because of financial and administrative constraints, ICS decided to phase in the program over four years. This allowed evaluators to use springs that had not received the treatment yet as the comparison group.

From the 200 eligible springs, 100 were randomly selected to receive the treatment in the first two years. The study found that spring protection reduced fecal water contamination by 66 percent and child diarrhea among users of the springs by 25 percent.

Source: Kremer and others 2011.

Box 4.6: Randomized Assignment of Information about HIV Risks to Curb Teen Pregnancy in Kenya

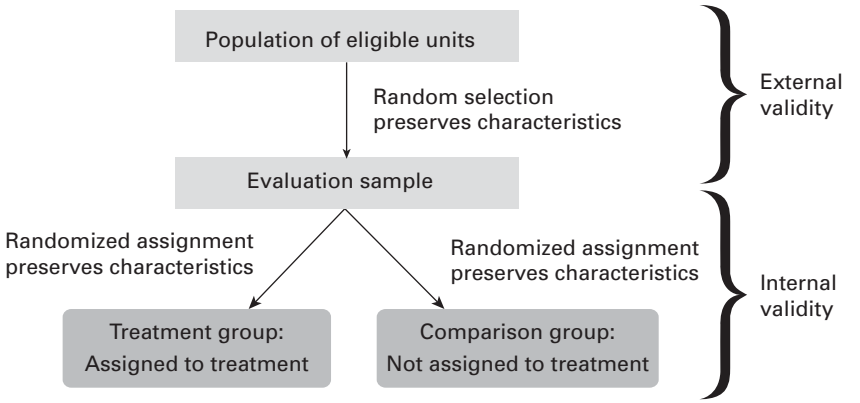
In a randomized experiment in western Kenya, Dupas (2011) tested the effectiveness of two different HIV/AIDS education treatments in reducing unsafe sexual behavior among teens. The first treatment involved teacher training in the national HIV/AIDS curriculum, which focused on risk aversion and encouraged abstinence. The second treatment, the Relative Risk Information Campaign, aimed to reduce sex between older men and younger girls by providing information on HIV rates disaggregated by age and gender.

The study took place in two rural districts in Kenya, with 328 primary schools in the sample. The researchers randomly assigned 163 schools to receive the first treatment, stratified by location, test scores, and student gender ratio. Seventy-one schools were then randomly assigned to the second treatment, stratifying for participation in the first treatment. This produced four groups of schools: schools receiving treatment one, schools receiving treatment two, schools receiving both treatments, and schools receiving neither treatment.

The randomized assignment of schools ensured there would be no systematic difference in the information students were exposed to before the program started. A year after the program ended, Dupas found that the Relative Risk Information Campaign led to a 28 percent decrease in the likelihood that a girl would be pregnant. In contrast, schools that received only treatment one showed no effect on teenage pregnancy.

Source: Dupas 2011.

Figure 4.2 Random Sampling and Randomized Assignment of Treatment



that is statistically equivalent to the treatment group at baseline, before the program starts. Once the program starts, the comparison group is exposed to the same set of external factors as the treatment group over time; the only exception is the program. Therefore, if any differences in outcomes appear between the treatment and comparison groups, they can only be due to the existence of the program in the treatment group. The internal validity of an impact evaluation is ensured through the process of *randomized assignment of treatment*.

Key Concept

An evaluation is internally valid if it provides an accurate estimate of the counterfactual through a valid comparison group.

External validity means that the evaluation *sample* accurately represents the population of eligible units. The results of the evaluation can then be generalized to the population of eligible units. We use *random sampling* to ensure that the evaluation sample accurately reflects the population of eligible units so that impacts identified in the evaluation sample can be extrapolated to the population.

Key Concept

An evaluation is externally valid if the evaluation sample accurately represents the population of eligible units. The results of the evaluation can then be generalized to the population of eligible units.

Note that we have used a randomization process for two different purposes: *random selection* of a sample (for external validity), and *randomized assignment* of treatment as an impact evaluation method (for internal validity). An impact evaluation can produce internally valid estimates of impact through randomized assignment of treatment; however, if the evaluation is performed on a nonrandom sample of the population, the estimated impacts may not be generalizable to the population of eligible units. Conversely, if the evaluation uses a random sample of the population of eligible units, but treatment is not assigned in a randomized way, then the sample would be representative, but the comparison group may not be valid, thus jeopardizing internal validity. In some contexts programs may face constraints that demand a trade-off between internal

and external validity. Take the program discussed previously that targets the bottom 20 percent of households based on income. If this program incorporates all households below the 15th percentile, but conducts a randomized assignment impact evaluation among a random sample of households in the 15th to 25th percentiles, this evaluation will have internal validity thanks to the randomized assignment: that is, we will know the true impact for the subset of households in the 15th to 25th percentiles. But external validity of the impact evaluation will be limited, since results cannot be extrapolated directly to the entire population of beneficiaries: in particular, to households below the 15th percentile.

When Can Randomized Assignment Be Used?

Randomized assignment can be used as a program allocation rule in one of two specific scenarios:

1. *When the eligible population is greater than the number of program spaces available.* When the demand for a program exceeds the supply, a lottery can be used to select the treatment group within the eligible population. In this context, every unit in the population receives the same chance (or a known chance greater than zero and less than one) of being selected for the program. The group that wins the lottery is the treatment group, and the rest of the population that is not offered the program is the comparison group. As long as a constraint exists that prevents scaling the program up to the entire population, the comparison groups can be maintained to measure the short-term, medium-term, and long-term impacts of the program. In this context, no ethical dilemma arises from holding a comparison group indefinitely, since a subset of the population will necessarily be left out of the program because of capacity constraints.

   As an example, suppose the ministry of education wants to provide school libraries to public schools throughout the country, but the ministry of finance budgets only enough funds to cover one-third of them. If the ministry of education wants each public school to have an equal chance of receiving a library, it would run a lottery in which each school has the same chance (1 in 3) of being selected. Schools that win the lottery receive a new library and constitute the treatment group, and the remaining two-thirds of public schools in the country are not offered the library and serve as the comparison group. Unless additional funds are allocated to the library program, a group of schools will remain that do not have funding for

libraries, and they can be used as a comparison group to measure the counterfactual.

2. *When a program needs to be gradually phased in until it covers the entire eligible population.* When a program is phased in, randomization of the order in which participants receive the program gives each eligible unit the same chance of receiving treatment in the first phase or in a later phase of the program. As long as the last group has not yet been phased into the program, it serves as a valid comparison group from which the counterfactual for the groups that have already been phased in can be estimated. This setup can also allow for the evaluation to pick up the effects of differential *exposure to treatment*: that is, the effect of receiving a program for more or less time.

   For example, suppose that the ministry of health wants to train all 15,000 nurses in the country to use a new health care protocol but needs three years to train them all. In the context of an impact evaluation, the ministry could randomly assign one-third of the nurses to receive training in the first year, one-third to receive training in the second year, and one-third to receive training in the third year. To evaluate the effect of the training program one year after its implementation, the group of nurses trained in year 1 would constitute the treatment group, and the group of nurses randomly assigned to training in year 3 would be the comparison group, since they would not yet have received the training.

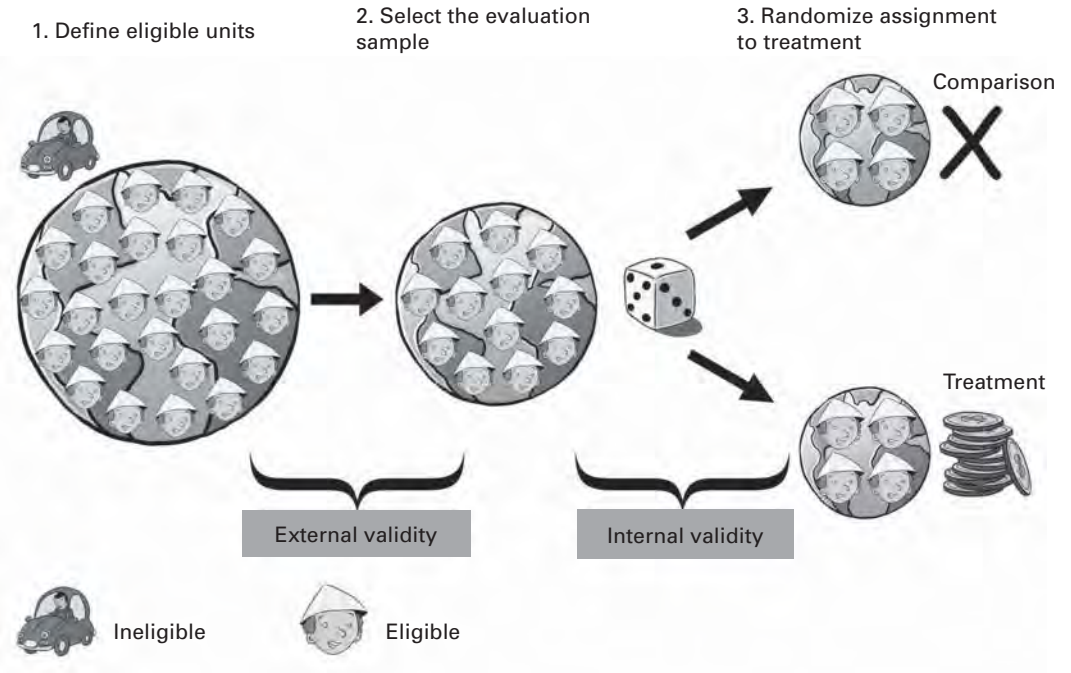
How Do You Randomly Assign Treatment?

Now that we have discussed what randomized assignment does and why it produces a good comparison group, we will turn to the steps to successfully assign treatment in a randomized way. Figure 4.3 illustrates this process.

Step 1 is to define the units that are eligible for the program. Remember that depending on the particular program, a unit could be a person, a health center, a school, a business, or even an entire village or municipality. The population of eligible units consists of those for which you are interested in knowing the impact of your program. For example, if you are implementing a training program for primary school teachers in rural areas, then primary school teachers in urban areas or secondary school teachers would not belong to your population of eligible units.

Once you have determined the population of eligible units, it will be necessary to compare the size of the group with the number of observations

Figure 4.3 Steps in Randomized Assignment to Treatment



required for the evaluation. The size of the evaluation sample is determined through power calculations and is based on the types of questions you would like answered (see chapter 15). If the eligible population is small, all of the eligible units may need to be included in the evaluation. Alternatively, if there are more eligible units than are required for the evaluation, then step 2 is to select a sample of units from the population to be included in the evaluation sample.

This second step is done mainly to limit data collection costs. If it is found that data from existing monitoring systems can be used for the evaluation, and that those systems cover the full population of eligible units, then you may not need to draw a separate evaluation sample. However, imagine an evaluation in which the population of eligible units includes tens of thousands of teachers in every school in the country, and you need to collect detailed information on teacher pedagogical knowledge and practice. Interviewing and assessing every teacher in the country could be prohibitively costly and logistically infeasible. Based on your power calculations, you might determine that to answer your evaluation question, it is sufficient to take a sample of 1,000 teachers distributed over

200 schools. As long as the sample of teachers is representative of the whole population of teachers, any results found in the evaluation will be externally valid and can be generalized to the rest of the teachers in the country. Collecting data on this sample of 1,000 teachers in 200 schools will be much cheaper than collecting data on every teacher in all schools in the country.

Step 3 is to form the treatment and comparison groups from the units in the evaluation sample through randomized assignment. In cases where randomized assignment needs to be done in a public forum, say on television, you may need to use a simple hands-on technique such as flipping a coin or picking names out of a hat. The following examples assume that the unit of randomization is an individual person, but the same logic applies to randomizing more aggregated units of observation such as schools, businesses, or communities:

1. If you want to assign 50 percent of individuals to the treatment group and 50 percent to the comparison group, flip a coin for each person. You must decide in advance whether heads or tails on the coin will assign a person to the treatment group.

2. If you want to assign one-third of the evaluation sample to the treatment group, you can roll a die for each person. First, you must decide on a rule. For example, a thrown die that shows a 1 or a 2 could mean an assignment to the treatment group, whereas a 3, 4, 5, or 6 would mean an assignment to the comparison group. You would roll the die once for each person in the evaluation sample and assign them based on the number that comes up.

3. Write the names of all of the people on pieces of paper of identical size and shape. Fold the papers so that the names cannot be seen, and mix them thoroughly in a hat or some other container. Before you start drawing, decide on your rule: that is, how many pieces of paper you will draw and that drawing a name means assigning that person to the treatment group. Once the rule is clear, ask someone in the crowd (someone unbiased, such as a child) to draw out as many pieces of paper as you need participants in the treatment group.

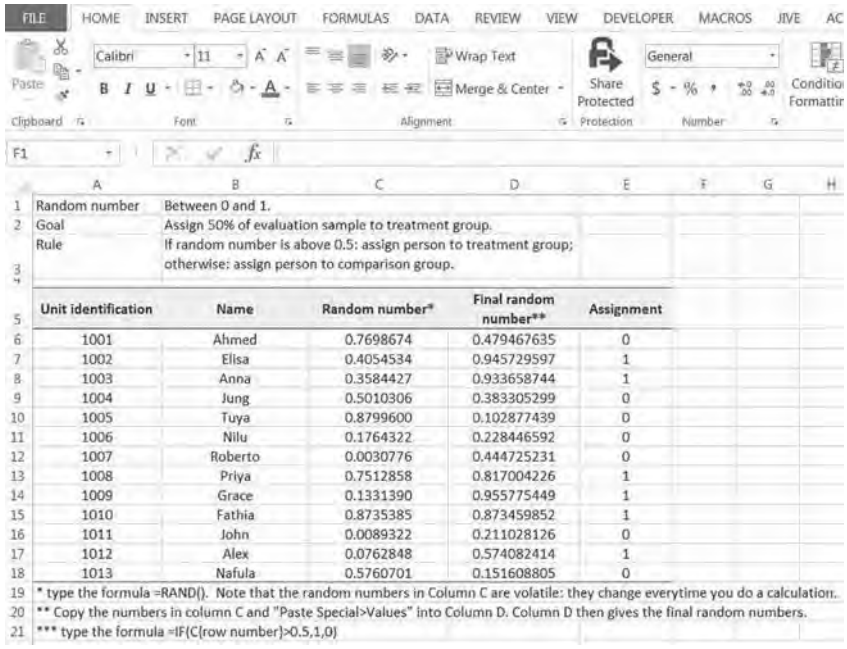
If you need to assign many units (say, over 100), using simple approaches such as these will take too much time, and you will need to use an automated process. To do this, you must first decide on a rule for how to assign participants based on random numbers. For example, if you need to assign 40 out of 100 units from the evaluation sample to the treatment group, you may decide to assign those 40 units with the highest random numbers to the treatment group and the rest to the comparison group. To implement

the randomized assignment, you will assign a random number to each unit in the evaluation sample, using a spreadsheet's random number generator, or specialized statistical software (figure 4.4), and use your previously chosen rule to form the treatment and comparison groups. It is important to decide on the rule before you generate the random numbers; otherwise, you may be tempted to decide on a rule based on the random numbers you see, and that would invalidate the randomized assignment.

The logic behind the automated process is no different from randomized assignment based on a coin toss or picking names out of a hat: it is a mechanism that randomly determines whether each unit is in the treatment or the comparison group.

Whether you use a public lottery, a roll of dice, or computer-generated random numbers, it is important to document the process to ensure that it is transparent. That means, first, that the assignment rule must be decided in advance and communicated to members of the public. Second, you must stick to the rule once you draw the random numbers. Third, you must be able to show that the process was really random. In the cases of lotteries and throwing dice, you could videotape the process; computer-based assignment of random numbers requires that you provide a log of your computations, so that the process can be replicated by auditors.[5]

Figure 4.4 Using a Spreadsheet to Randomize Assignment to Treatment



	A	B	C	D	E	F	G	H
1	Random number	Between 0 and 1.						
2	Goal	Assign 50% of evaluation sample to treatment group.						
3	Rule	If random number is above 0.5: assign person to treatment group; otherwise: assign person to comparison group.						
5	**Unit identification**	**Name**	**Random number***	**Final random number****	**Assignment**			
6	1001	Ahmed	0.7698674	0.479467635	0			
7	1002	Elisa	0.4054534	0.945729597	1			
8	1003	Anna	0.3584427	0.933658744	1			
9	1004	Jung	0.5010306	0.383305299	0			
10	1005	Tuya	0.8799600	0.102877439	0			
11	1006	Nilu	0.1764322	0.228446592	0			
12	1007	Roberto	0.0030776	0.444725231	0			
13	1008	Priya	0.7512858	0.817004226	1			
14	1009	Grace	0.1331390	0.955775449	1			
15	1010	Fathia	0.8735385	0.873459852	1			
16	1011	John	0.0089322	0.211028126	0			
17	1012	Alex	0.0762848	0.574082414	1			
18	1013	Nafula	0.5760701	0.151608805	0			
19	* type the formula =RAND(). Note that the random numbers in Column C are volatile: they change everytime you do a calculation.							
20	** Copy the numbers in column C and "Paste Special>Values" into Column D. Column D then gives the final random numbers.							
21	*** type the formula =IF(C{row number}>0.5,1,0)							

At What Level Do You Perform Randomized Assignment?

Randomized assignment can be done at various levels, including the individual, household, business, community, or region. In general, the level at which units are randomly assigned to treatment and comparison groups will be greatly affected by where and how the program is being implemented. For example, if a health program is being implemented at the health clinic level, you would first select a random sample of health clinics and then randomly assign some of them to the treatment group and others to the comparison group.

When the level of the randomized assignment is higher or more aggregate, such as at the level of regions or provinces in a country, it can become difficult to perform an impact evaluation because the number of regions or provinces in most countries is not sufficiently large to yield balanced treatment and comparison groups. For example, if a country has only six provinces, then the treatment and comparison groups would only have three provinces each, which is insufficient to ensure that the baseline characteristics of the treatment and comparison groups are balanced. Furthermore, for randomized assignment to yield unbiased estimates of impact, it is important to ensure that time-bound external factors (such as the weather or local election cycles) are on average the same in the treatment and comparison groups. As the level of assignment increases, it becomes increasingly unlikely that these factors will be balanced across treatment and comparison groups. For example, rainfall is a time-bound external factor because it varies systematically from one year to another. In an evaluation in the agriculture sector, we would want to ensure that droughts affect the treatment and comparison provinces equally. With only three provinces in the treatment and comparison groups, it would be easy for this balance to be lost. On the other hand, if the unit of assignment were lowered to the subprovince level such as a municipality, it is more likely that rainfall will be balanced between treatment and comparison groups over time.

On the other hand, as the level of randomized assignment gets lower—for example, down to the individual or household level—the chances increase that the comparison group will be inadvertently affected by the program. Two particular types of risks to consider when choosing the level of assignment are spillovers and imperfect compliance. *Spillovers* occur when the treatment group directly or indirectly affects outcomes in the comparison group (or vice versa). *Imperfect compliance* occurs when some members of the comparison group participate in the program, or some members of the treatment group do not (see further discussion of these concepts in chapter 9).

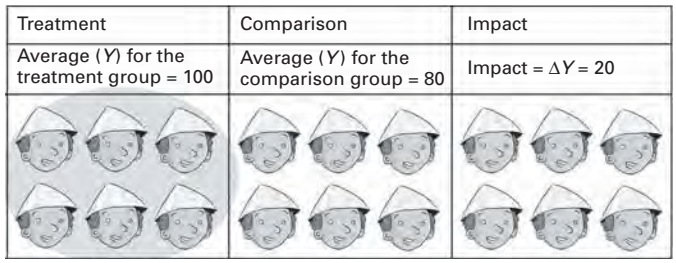
By carefully considering the level of randomized assignment, the risk of spillovers and imperfect compliance can be minimized. Individuals can be assigned in groups or clusters such as students in a school or households in a community, to minimize information flows and contacts between individuals in the treatment and comparison groups. To reduce imperfect compliance, the level of assignment should also be chosen in accordance with the program's capacity for maintaining a clear distinction between treatment and comparison groups throughout the intervention. If the program includes activities at the community level, it may be difficult to avoid exposing all individuals from that community to the program.

A well-known example of spillovers is the provision of deworming medicine to children. If households in the treatment group are located close to a household in the comparison group, then children in the comparison households may be positively affected by a spillover from the treatment because their chances of contracting worms from neighbors will be reduced (Kremer and Miguel 2004). To isolate the program impact, treatment and comparison households need to be located sufficiently far from one another to avoid such spillovers. Yet as the distance between households increases, it will become more costly both to implement the program and to administer surveys. As a rule of thumb, if spillovers can be reasonably ruled out, it is best to perform randomized assignment of the treatment at the lowest possible level of program implementation; that will ensure that the number of units in the treatment and comparison groups is as large as possible.

Estimating Impact under Randomized Assignment

Once you have selected a random evaluation sample and assigned treatment in a randomized fashion, it is quite straightforward to estimate the impact of the program. After the program has run for some time, outcomes for both the treatment and comparison units will need to be measured. The impact of the program is simply the difference between the average outcome (Y) for the treatment group and the average outcome (Y) for the comparison group. For instance, in the generic example in figure 4.5, the average outcome for the treatment group is 100, and the average outcome for the comparison group is 80; thus the impact of the program is 20. For now we are assuming that all units in the treatment group are effectively treated and no units of the comparison group are treated. In our example of the teacher training program, all teachers assigned to the treatment group receive the training and none of the

Figure 4.5 Estimating Impact under Randomized Assignment



comparison teachers do. In chapter 5, we discuss the (more realistic) scenario where compliance is incomplete: that is, where less than 100 percent of the units in the treatment group actually participate in the intervention or some comparison units gain access to the program. In this case, an unbiased estimate of program impact can still be obtained through randomized assignment, though the interpretation of the results will vary.

Checklist: Randomized Assignment

Randomized assignment is the most robust method for estimating counterfactuals; it is considered the gold standard of impact evaluation. Some basic tests should still be considered to assess the validity of this evaluation strategy in a given context.

- ✓ Are the baseline characteristics balanced? Compare the baseline characteristics of the treatment group and the comparison group.[6]
- ✓ Has any noncompliance with the assignment occurred? Check whether all eligible units have received the treatment and that no ineligible units have received the treatment. If noncompliance has occurred, you will need to use the instrumental variable method (see chapter 5).
- ✓ Are the numbers of units in the treatment and comparison groups sufficiently large? If not, you may want to combine randomized assignment with difference-in-differences (see chapter 7).
- ✓ Is there any reason to believe that outcomes for some units may somehow depend on the assignment of other units? Could there be an impact of the treatment on units in the comparison group (see chapter 9)?

Evaluating the Impact of HISP: Randomized Assignment

Let us now return to the example of the Health Insurance Subsidy Program (HISP) and check what randomized assignment means in this context. Recall that you are trying to estimate the impact of the program from a pilot that involves 100 treatment villages.

Having conducted two impact assessments using potentially biased estimators of the counterfactual in chapter 3 (with conflicting policy recommendations), you decide to go back to the drawing board to rethink how to obtain a more precise estimate of the counterfactual. After further deliberations with your evaluation team, you are convinced that constructing a valid estimate of the counterfactual will require identifying a group of villages that are as similar as possible to the 100 treatment villages in all respects, except that one group took part in HISP and the other did not. Because HISP was rolled out as a pilot, and the 100 treatment villages were selected randomly from among all of the rural villages in the country, you note that the treatment villages should, on average, have the same characteristics as the untreated rural villages in the country. The counterfactual can therefore be estimated in a valid way by measuring the health expenditures of eligible households in rural villages that did not take part in the program.

Luckily, at the time of the baseline and follow-up surveys, the survey firm collected data on an additional 100 rural villages that were not offered the program. Those 100 villages were also randomly selected from the population of rural villages in the country. Thus the way that the two groups of villages were chosen ensures that they have statistically identical characteristics, except that the 100 treatment villages received HISP and the 100 comparison villages did not. Randomized assignment of the treatment has occurred.

Given randomized assignment of treatment, you are quite confident that no external factors other than HISP would explain any differences in outcomes between the treatment and comparison villages. To validate this assumption, you test whether eligible households in the treatment and comparison villages have similar characteristics at baseline, as shown in table 4.1.

You observe that the average characteristics of households in the treatment and comparison villages are in fact very similar. The only statistically significant differences are for the number of years of education of the head of household and distance to hospital, and those differences are small (only 0.16 years, or less than 6 percent of the

Table 4.1 Evaluating HISP: Balance between Treatment and Comparison Villages at Baseline

Household characteristics	Treatment villages (n = 2964)	Comparison villages (n = 2664)	Difference	t-stat
Health expenditures (US$ yearly per capita)	14.49	14.57	−0.08	−0.73
Head of household's age (years)	41.66	42.29	−0.64	−1.69
Spouse's age (years)	36.84	36.88	0.04	0.12
Head of household's education (years)	2.97	2.81	0.16*	2.30
Spouse's education (years)	2.70	2.67	0.03	0.43
Head of household is female = 1	0.07	0.08	−0.01	−0.58
Indigenous = 1	0.43	0.42	0.01	0.69
Number of household members	5.77	5.71	0.06	1.12
Has dirt floor = 1	0.72	0.73	−0.01	−1.09
Has bathroom = 1	0.57	0.56	0.01	1.04
Hectares of land	1.68	1.72	−0.04	−0.57
Distance to hospital (km)	109.20	106.29	2.91*	2.57

Note: Significance level: ** = 1 percent.

Table 4.2 Evaluating HISP: Randomized Assignment with Comparison of Means

	Treatment villages	Comparison villages	Difference	t-stat
Household health expenditures at baseline (US$)	14.49	14.57	−0.08	−0.73
Household health expenditures at follow-up (US$)	7.84	17.98	−10.14**	−49.15

Note: Significance level: ** = 1 percent

comparison group's average years of education, and 2.91 kilometers, or less than 3 percent of the comparison group's average distance to a hospital). Even with a randomized experiment on a large sample, a small number of differences can be expected because of chance and

the properties of the statistical test. In fact, using standard significance levels of 5 percent we could expect differences in about 5 percent of characteristics to be statistically significant, though we would not expect the magnitude of these differences to be large.

With the validity of the comparison group established, you can now estimate the counterfactual as the average health expenditures of eligible households in the 100 comparison villages. Table 4.2 shows the average household health expenditures for eligible households in the treatment and comparison villages. You note that at baseline, the average household health expenditures in the treatment and comparison groups are not statistically different, as should be expected under randomized assignment.

Given that you now have a valid comparison group, you can find the impact of the HISP simply by taking the difference between the average out-of-pocket health expenditures of households in the treatment villages and randomly assigned comparison villages in the follow-up period. The impact is a reduction of US$10.14 over two years. Replicating this result through a linear regression analysis yields the same result, as shown in table 4.3. Finally, you run a multivariate regression analysis that controls for some other observable characteristics of the sample households, and you find that the program has reduced the expenditures of the enrolled households by US$10.01 over two years, which is nearly identical to the linear regression result.

With randomized assignment, we can be confident that no factors are present that are systematically different between the treatment and comparison groups that might also explain the difference in health expenditures. Both sets of villages started off with very similar average characteristics and have been exposed to the same set of national policies and programs during the two years of treatment. Thus the only plausible reason that poor households in treatment communities have lower expenditures than households in comparison villages is that the first group received the health insurance program and the other group did not.

Table 4.3 Evaluating HISP: Randomized Assignment with Regression Analysis

	Linear regression	Multivariate linear regression
Estimated impact on household health expenditures	−10.14**	−10.01**
	(0.39)	(0.34)

Note: Standard errors are in parentheses. Significance level: ** = 1 percent.

HISP Question 3

A. Why is the impact estimate derived using a multivariate linear regression basically unchanged when controlling for other factors, compared to the simple linear regression and comparison of means?

B. Based on the impact estimated with the randomized assignment method, should the HISP be scaled up nationally?

Additional Resources

- For accompanying material to this chapter and hyperlinks to additional resources, please see the Impact Evaluation in Practice website (www.worldbank.org/ieinpractice).
- For additional resources on randomized assignment impact evaluations, see the Inter-American Development Bank Evaluation Portal (www.iadb.org/evaluationhub).
- For a complete overview of randomized assignment impact evaluations, see the following book and accompanying website:
  - Glennerster, Rachel, and Kudzai Takavarasha. 2013. *Running Randomized Evaluations: A Practical Guide*. Princeton, NJ: Princeton University Press (http://runningres.com/).
- For a detailed discussion on achieving balance between treatment and comparison groups through randomized assignment, see the following:
  - Bruhn, Miriam, and David McKenzie. 2009. "In Pursuit of Balance: Randomization in Practice in Development Field Experiments." *American Economic Journal: Applied Economics* 1 (4): 200–32.
- For a randomized assignment ceremony for an evaluation in Cameroon, see the World Bank Impact Evaluation Toolkit, Module 3 (www.worldbank.org/health/impactevaluationtoolkit).

Notes

1. Randomized assignment of treatment is also commonly referred to as *randomized control trials, randomized evaluations, experimental evaluations, and social experiments,* among other terms. Strictly speaking, an experiment need not identify impacts through randomized assignment, but evaluators typically use the term *experiment* only when the evaluation uses randomized assignment.
2. Note that this probability does not necessarily mean a 50-50 chance of winning the lottery. In practice, most randomized assignment evaluations will give each eligible unit a probability of selection that is determined so that the number of

winners (treatments) equals the total available number of benefits. For example, if a program has enough funding to serve only 1,000 communities out of a population of 10,000 eligible communities, then each community will be given a 1 in 10 chance of being selected for treatment. Statistical power (a concept discussed in more detail in chapter 15) will be maximized when the evaluation sample is divided equally between the treatment and comparison groups. In the example here, for a total sample size of 2,000 communities, statistical power will be maximized by sampling all 1,000 treatment communities and a subsample of 1,000 comparison communities, rather than by taking a simple random sample of 20 percent of the original 10,000 eligible communities (which would produce an evaluation sample of roughly 200 treatment communities and 1,800 comparison communities).

3. For example, housing programs that provide subsidized homes routinely use lotteries to select program participants. Many charter schools in the United States use lotteries to select which applicants are granted admission.
4. In addition to creating groups that have similar average characteristics, randomized assignment also creates groups that have similar distributions.
5. Most software programs allow you to set a *seed number* to make the results of the randomized assignment fully transparent and replicable.
6. As mentioned, for statistical reasons, not all observed characteristics must be similar in the treatment and comparison groups for randomization to be successful. Even when the characteristics of the two groups are truly equal, one can expect that 5 percent of the characteristics will show up with a statistically significant difference when a 95 percent confidence level is used for the test. Of particular concern are variables where the difference between treatment and comparison groups is large.

References

Bertrand, Marianne, Bruno Crépon, Alicia Marguerie, and Patrick Premand. 2016. "Impacts à Court et Moyen Terme sur les Jeunes des Travaux à Haute Intensité de Main d'oeuvre (THIMO): Résultats de l'évaluation d'impact de la composante THIMO du Projet Emploi Jeunes et Développement des compétence (PEJEDEC) en Côte d'Ivoire." Washington, DC: Banque Mondiale et Abidjan, BCP-Emploi.

Blattman, Christopher, Nathan Fiala, and Sebastian Martinez. 2014. "Generating Skilled Self-Employment in Developing Countries: Experimental Evidence from Uganda." *Quarterly Journal of Economics* 129 (2): 697–752. doi: 10.1093/qje/qjt057.

Bruhn, Miriam, and David McKenzie. 2009. "In Pursuit of Balance: Randomization in Practice in Development Field Experiments." *American Economic Journal: Applied Economics* 1 (4): 200–232.

Dupas, Pascaline. 2011. "Do Teenagers Respond to HIV Risk Information? Evidence from a Field Experiment in Kenya." *American Economic Journal: Applied Economics* 3 (1): 1–34.

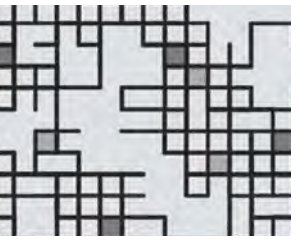
Glennerster, Rachel, and Kudzai Takavarasha. 2013. *Running Randomized Evaluations: A Practical Guide*. Princeton, NJ: Princeton University Press.

Kremer, Michael, Jessica Leino, Edward Miguel, and Alix Peterson Zwane. 2011. "Spring Cleaning: Rural Water Impacts, Valuation, and Property Rights Institutions." *Quarterly Journal of Economics* 126: 145–205.

Kremer, Michael, and Edward Miguel. 2004. "Worms: Identifying Impacts on Education and Health in the Presence of Treatment Externalities." *Econometrica* 72 (1): 159–217.

Premand, Patrick, Oumar Barry, and Marc Smitz. 2016. "Transferts monétaires, valeur ajoutée de mesures d'accompagnement comportemental, et développement de la petite enfance au Niger. Rapport descriptif de l'évaluation d'impact à court terme du Projet Filets Sociaux." Washington, DC: Banque Mondiale.

Schultz, Paul. 2004. "School Subsidies for the Poor: Evaluating the Mexican Progresa Poverty Program." *Journal of Development Economics* 74 (1): 199–250.



CHAPTER 5

Instrumental Variables

Evaluating Programs When Not Everyone Complies with Their Assignment

In the discussion of randomized assignment in chapter 4, we assumed that the program administrator has the power to assign units to treatment and comparison groups, with those assigned to the treatment taking the program and those assigned to the comparison group not taking the program. In other words, units that are assigned to the treatment and comparison groups comply with their assignment. Full compliance is more frequently attained in laboratory settings or medical trials, where the researcher can carefully make sure, first, that all subjects in the treatment group take a given treatment, and second, that none of the subjects in the comparison group take it.[1] More generally in chapter 4, we assumed that programs are able to determine who the potential participants are, excluding some and ensuring that others participate.

However, in real-world social programs, it might be unrealistic to think that the program administrator will be able to ensure full compliance with the group assignment. Yet many programs allow potential participants to choose to enroll and thus are not able to exclude potential participants who want to enroll. In addition, some programs have a budget that is big enough to supply the program to the entire eligible population immediately, so that randomly assigning people to treatment and comparison groups and

excluding potential participants for the sake of an evaluation would not be ethical. We therefore need an alternative way to evaluate the impact of these kinds of programs.

Key Concept

The instrumental variable method relies on some external source of variation to determine treatment status. An instrumental variable influences the likelihood of participating in a program, but is outside of the participant's control and is unrelated to the participant's characteristics.

A method called *instrumental variables* (IV) can help us evaluate programs with imperfect compliance, voluntary enrollment, or universal coverage. Generally, to estimate impacts, the IV method relies on some external source of variation to determine treatment status. The method has wide-ranging applications beyond impact evaluation. Intuitively, we can think of an IV as something outside the control of the individual that influences her likelihood of participating in a program, but is otherwise not associated with her characteristics.

In this chapter, we discuss how this external variation, or IV, can be generated by the rules of program operation that are under the control of program implementers or evaluation teams. To produce valid impact estimates, this external source of variation must satisfy a number of conditions, which we will discuss in detail in this chapter. It turns out that randomized assignment of treatment, as discussed in chapter 4, is a very good instrument, satisfying the necessary conditions. We will use the IV method in two common applications of impact evaluation. First, we will use it as an extension of the randomized assignment method when not all units comply with their group assignments. Second, we will use it to design randomized promotion of treatment, an evaluation method that can work for some programs that offer voluntary enrollment or universal coverage. Box 5.1 illustrates a creative use of the IV method.

Types of Impact Estimates

An impact evaluation always estimates the impact of a program by comparing the outcomes for a treatment group with the estimate of the counterfactual obtained from a comparison group. In chapter 4, we assumed *full compliance* with treatment: that is, all units to whom a program has been offered actually enroll, and none of the comparison units receive the program. In this scenario, we estimate the *average treatment effect* (ATE) for the population.

In the evaluation of real-world programs where potential participants can decide whether to enroll or not, full compliance is less common than in settings such as laboratory experiments. In practice, programs typically offer the opportunity of treatment to a particular group, and some units participate while others do not. In this case, without full compliance, impact evaluations can estimate the effect of *offering* a program or the effect of *participating* in the program.

Box 5.1: Using Instrumental Variables to Evaluate the Impact of *Sesame Street* on School Readiness

The television show *Sesame Street*, a program aimed at preparing preschool-aged children for primary school, quickly gained critical acclaim and popularity after first airing in 1969. It has since been watched by millions of children. In 2015, Kearney and Levine sought to evaluate the long-term impacts of the program in a retrospective evaluation carried out in the United States. Taking advantage of limitations in television broadcasting technology in the early years of the show, the researchers used an instrumental variables approach.

In the first few years the show was not accessible to all households. It was only broadcast on ultra-high frequency (UHF) channels. Only about two-thirds of the U.S. population lived in areas where the show was accessible. Thus, Kearney and Levine (2015) used households' distance to the closest television tower that transmitted UHF as an instrument for participation in the program. The researchers argue that since television towers were built in locations chosen by the government—all before *Sesame Street* was ever broadcast—the variable would not be related to household characteristics or changes in the outcome.

The evaluation found positive results on school readiness for preschool-aged children. In areas where there was UHF television reception when the show began, children were more likely to advance through primary school at the appropriate age. This effect was notable for African-American and non-Hispanic children, boys, and children in economically disadvantaged areas.

Source: Kearney and Levine 2015.

In the absence of full compliance in the treatment group, the estimated impact Δ is called the *intention-to-treat* (ITT) when comparing groups to which the program has randomly been *offered* (in the treatment group) or not (in the comparison group)—regardless of whether or not those in the treatment group actually enroll in the program. The ITT is a weighted average of the outcomes of participants and nonparticipants in the treatment group compared with the average outcome of the comparison group. The ITT is important for those cases in which we are trying to determine the average impact of offering a program, and enrollment in the treatment group is voluntary. By contrast, we might also be interested in knowing the impact of a program for the group of individuals who are offered the program and actually participate. This estimated impact is called the *treatment-on-the-treated* (TOT). The ITT and TOT will be the same when there is full compliance. We will return to the difference between the ITT and TOT in future sections, but start with an example to illustrate these concepts.

Key Concept

Intention-to-treat (ITT) estimates the difference in outcomes between the units assigned to the treatment group and the units assigned to the comparison group, irrespective of whether the units assigned to the treatment group actually receive the treatment.

Key Concept

Treatment-on-the-treated (TOT) estimates the difference in outcomes between the units that actually receive the treatment and the comparison group.

Consider the Health Insurance Subsidy Program (HISP), discussed in previous chapters. Because of operational considerations and to minimize spillovers, the unit of treatment assignment chosen by the government is

the village. Households in a treatment village (the villages where the health insurance program is being offered) can sign up for a health insurance subsidy voluntarily, while households in comparison communities cannot. Even though all households in treatment villages are eligible to enroll in the health insurance program, some fraction of households—say, 10 percent—may decide not to do so (perhaps because they already have insurance through their jobs, because they are healthy and do not anticipate the need for health care, or because of any other myriad reasons).

In this scenario, 90 percent of households in the treatment village decide to enroll in the program and actually receive the services that the program provides. The ITT estimate would be obtained by comparing the average outcome for all households that were offered the program—that is, for 100 percent of the households in treatment villages—with the average outcome in the comparison villages (where no households have enrolled). By contrast, the TOT can be thought of as the estimated impact for the 90 percent of households in treatment villages that enrolled in the program. It is important to note that since individuals who participate in a program when offered may differ from individuals who are offered the program but opt out, the TOT impact is not necessarily the same as the impact we would obtain for the 10 percent of households in the treatment villages that did not enroll, should they become enrolled. As such, local treatment effects cannot be extrapolated directly from one group to another.

Imperfect Compliance

As discussed, in real-world social programs, full compliance with a program's selection criteria (and hence adherence to treatment or comparison status) is desirable, and policy makers and evaluation teams alike usually strive to come as close to that ideal as possible. In practice, however, strict 100 percent compliance to treatment and comparison assignments may not occur, despite the best efforts of the program implementer and the evaluation team. We will now work through the different cases that can occur and discuss implications for the evaluation methods that can be used. We stress up front that the best solution to imperfect compliance is to avoid it in the first place. In this sense, program managers and policy makers should strive to keep compliance as high as possible in the treatment group and as low as possible in the comparison group.

Say you are trying to evaluate a teacher-training program, in which 2,000 teachers are eligible to participate in a pilot training. The teachers have been randomly assigned to one of two groups: 1,000 teachers are assigned to the treatment group and 1,000 teachers are assigned to the comparison group.

When all teachers in the treatment group receive training, and none in the comparison group have, we estimate the ATE by taking the difference in mean outcomes (say student test scores) between the two groups. This ATE is the average impact of the treatment on the 1,000 teachers, given that all teachers assigned to the treatment group actually attend the course, while none of the teachers assigned to the comparison group attend.

The first case of imperfect compliance occurs when some units assigned to the treatment group choose not to enroll or are otherwise left untreated. In the teacher-training example, some teachers assigned to the treatment group do not actually show up on the first day of the course. In this case, we cannot calculate the average treatment for the population of teachers because some teachers never enroll; therefore we can never calculate what their outcomes would have been with treatment. But we can estimate the average impact of the program on those teachers who actually take up or accept the treatment. We want to estimate the impact of the program on those teachers to whom treatment was assigned *and* who actually enrolled. This is the *TOT estimate*. In the teacher-training example, the TOT estimate provides the impact for teachers assigned to the treatment group who actually show up and receive the training.

The second case of imperfect compliance is when individuals assigned to the comparison group manage to participate in the program. Here the impacts cannot be directly estimated for the entire treatment group because some of their counterparts in the comparison group cannot be observed without treatment. The treated units in the comparison group were supposed to generate an estimate of the counterfactual for some units in the treatment group, but they receive the treatment; therefore there is no way of knowing what the program's impact would have been for this subset of individuals. In the teacher-training example, say that the most motivated teachers in the comparison group manage to attend the course somehow. In this case, the most motivated teachers in the treatment group would have no counterparts in the comparison group, and so it would not be possible to estimate the impact of the training for that segment of motivated teachers.

When there is noncompliance on either side, you should consider carefully what type of treatment effect you estimate and how to interpret them. A first option is to compute a straight comparison of the group originally assigned to treatment with the group originally assigned to comparison; this will yield the *ITT estimate*. The ITT compares those whom we intended to treat (those assigned to the treatment group) with those whom we intended not to treat (those assigned to the comparison group). If the noncompliance is only on the treatment side, this can be an interesting and relevant measure of impact because in any case most policy makers and program managers can only offer a program and cannot force the program on their target population.

In the teacher-training example, the government may want to know the average impact of the program for all assigned teachers, even if some of the teachers do not attend the course. This is because even if the government expands the program, there are likely to be teachers who will never attend. However, if there is noncompliance on the comparison side, the intention-to-treat estimate is not as insightful. In the case of the teacher training, since the comparison group of teachers includes teachers who are trained, the average outcome in the comparison group has been affected by treatment. Let's assume that the effect of teacher training on outcomes is positive. If the noncompliers in the comparison group are the most motivated teachers and they benefit the most from training, the average outcome for the comparison group will be biased upward (because the motivated teachers in the comparison group who got trained will increase the average outcome) and the ITT estimate will be biased downward (since it is the difference between the average outcomes in the treatment and comparison groups).

Under these circumstances of noncompliance, a second option is to estimate what is known as the *local average treatment effect* (LATE). LATE needs to be interpreted carefully, as it represents program effects for only a specific subgroup of the population. In particular, when there is noncompliance in both the treatment group and in the comparison group, the LATE is the impact on the subgroup of compliers. In the teacher-training example, if there is noncompliance in both the treatment and comparison group, then the LATE estimate is valid only for teachers in the treatment group who enrolled in the program and who would have not enrolled had they been assigned to the comparison group.

In the remainder of this section, we will explain how to estimate the LATE, and equally importantly, how to interpret the results. The LATE estimation principles apply when there is noncompliance in the treatment group, comparison group, or both simultaneously. The TOT is simply a LATE in the more specific case when there is noncompliance only in the treatment group. Therefore, the rest of this chapter focuses on how to estimate LATE.

Randomized Assignment of a Program and Final Take-Up

Imagine that you are evaluating the impact of a job-training program on individuals' wages. The program is randomly assigned at the individual level. The treatment group is assigned to the program, while the comparison group is not. Most likely, you will find three types of individuals in the population:

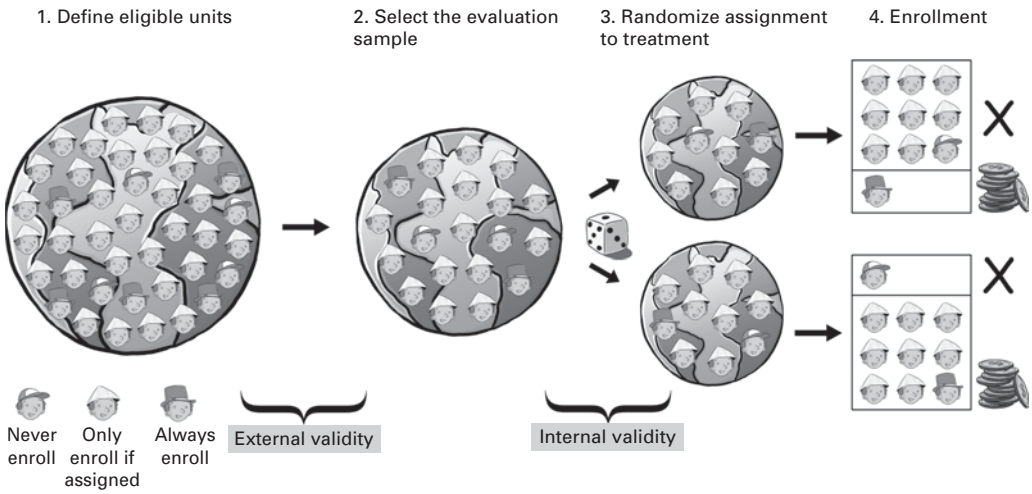
- *Enroll-if-assigned*. These are the individuals who comply with their assignment. If they are assigned to the treatment group (assigned to the program), they take it up, or enroll. If they are assigned to the comparison group (not assigned to the program), they do not enroll.

- *Never*. These are the individuals who never enroll in or take up the program, even if they are assigned to the treatment group. If assigned to the treatment group, these individuals will be *noncompliers*.
- *Always*. These are the individuals who will find a way to enroll in the program or take it up, even if they are assigned to the comparison group. If assigned to the comparison group, these individuals will be *noncompliers*.

In the context of the job-training program, the *Never* group might consist of unmotivated people who, even if assigned a place in the course, do not show up. Individuals in the *Always* group, in contrast, are so motivated that they find a way to enter the program even if they were originally assigned to the comparison group. The *Enroll-if-assigned* group comprises those who enroll in the course if they are assigned to it, but who do not seek to enroll if they are assigned to the comparison group.

Figure 5.1 presents the randomized assignment of the program and the final enrollment, or take-up, when *Enroll-if-assigned, Never,* and *Always* types are present. Say that the population comprises 80 percent *Enroll-if-assigned,* 10 percent *Never,* and 10 percent *Always.* If we take a random sample of the population for the evaluation sample, then the evaluation sample will also have approximately 80 percent *Enroll-if-assigned,* 10 percent *Never,* and 10 percent *Always.* Then if we randomly assign the

Figure 5.1 Randomized Assignment with Imperfect Compliance



evaluation sample to a treatment group and a comparison group, we should again have approximately 80 percent *Enroll-if-assigned,* 10 percent *Never,* and 10 percent *Always* in both groups. In the group that is assigned treatment, the *Enroll-if-assigned* and *Always* individuals will enroll, and only the *Never* group will stay away. In the comparison group, the *Always* will enroll, while the *Enroll-if-assigned* and *Never* groups will stay out. It is important to remember that while we know that these three types of individuals exist in the population, we can not necessarily distinguish an individual's type until we observe certain behaviors. In the treatment group, we will be able to identify the *Never* types when they fail to enroll, but we will not be able to distinguish the *Enroll-if-assigned* from the *Always,* since both types will enroll. In the comparison group, we will be able to identify the *Always* when they enroll, but we won't be able to distinguish between the *Enroll-if-assigned* and the *Never,* since both these types remain unenrolled.

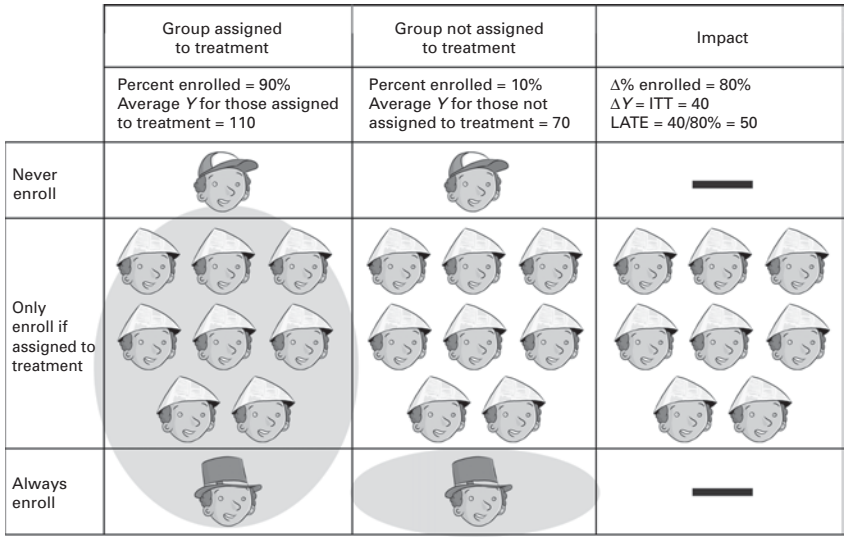
Estimating Impact under Randomized Assignment with Imperfect Compliance

Having established the difference between assigning a program and actual enrollment or take-up, we turn to estimating the LATE of the program. This estimation is done in two steps, which are illustrated in figure 5.2.[2]

To estimate program impacts under randomized assignment with imperfect compliance, we first estimate the ITT impact. Remember that this is just the straight difference in the outcome indicator (Y) for the group that we assigned to treatment and the same indicator for the group that we did not assign to treatment. For example, if the average wage (Y) for the treatment group is US$110, and the average wage for the comparison group is US$70, then the intention-to-treat estimate of the impact would be US$40 (US$110 minus US$70).

Second, we need to recover the LATE estimate for the *Enroll-if-assigned* group from the ITT estimate. To do that, we will need to identify where the US$40 difference came from. Let us proceed by elimination. First, we know that the difference cannot be caused by any differences between the people who never enroll (the *Nevers*) in the treatment and comparison groups. That's because the *Nevers* never enroll in the program, so for them, it makes no difference whether they are in the treatment group or in the comparison group. Second, we know that the US$40 difference cannot be caused by differences between the *Always* people in the treatment and comparison groups because the *Always* people always enroll in the program. For them, too, it makes no difference whether they

Figure 5.2 Estimating the Local Average Treatment Effect under Randomized Assignment with Imperfect Compliance



Note: Δ = causal impact; Y = outcome. The intention-to-treat (ITT) estimate is obtained by comparing outcomes for those assigned to the treatment group with those assigned to the comparison group, irrespective of actual enrollment. The local average treatment effect (LATE) estimate provides the impact of the program on those who enroll only if assigned to the program (*Enroll-if-assigned*). The LATE estimate does not provide the impact of the program on those who never enroll (the *Nevers*) or on those who always enroll (the *Always*).

are in the treatment group or the comparison group. Thus the difference in outcomes between the two groups must necessarily come from the effect of the program on the only group affected by their assignment to treatment or comparison: that is, the *Enroll-if-assigned* group. So if we can identify the *Enroll-if-assigned* in both groups, it will be easy to estimate the impact of the program on them.

In reality, although we know that these three types of individuals exist in the population, we cannot separate out unique individuals by whether they are *Enroll-if-assigned, Never,* or *Always*. In the group that was assigned treatment, we can identify the *Nevers* (because they have not enrolled), but we cannot differentiate between the *Always* and the *Enroll-if-assigned* (because both are enrolled). In the group that was not assigned treatment, we can identify the *Always* group (because they enroll in the program), but we cannot differentiate between the *Nevers* and the *Enroll-if-assigned*.

However, once we observe that 90 percent of the units in the group that was assigned treatment do enroll, we can deduce that 10 percent of the units in our population must be *Nevers* (that is, the fraction of individuals in the group assigned treatment who did not enroll). In addition, if we observe that 10 percent of units in the group not assigned treatment enroll, we know that 10 percent are *Always* (again, the fraction of individuals in our group that was not assigned treatment who did enroll). This leaves 80 percent of the units in the *Enroll-if-assigned* group. We know that the entire impact of US$40 came from a difference in enrollment for the 80 percent of the units in our sample who are *Enroll-if-assigned*. Now if 80 percent of the units are responsible for an average impact of US$40 for the entire group assigned treatment, then the impact on those 80 percent of *Enroll-if-assigned* must be 40/0.8, or US$50. Put another way, the impact of the program for the *Enroll-if-assigned* is US$50, but when this impact is spread across the entire group assigned treatment, the average effect is watered down by the 20 percent that was noncompliant with the original randomized assignment.

Remember that one of the basic issues with self-selection into programs is that you cannot always know why some people choose to participate and others do not. When we conduct an evaluation where units are randomly assigned to the program, but actual participation is voluntary or a way exists for units in the comparison group to get into the program, then we have a similar problem: we will not always understand the behavioral processes that determine whether an individual behaves like a *Never*, an *Always*, or an *Enroll-if-assigned*. However, provided that the noncompliance is not too large, randomized assignment still provides a powerful tool for estimating impact. The downside of randomized assignment with imperfect compliance is that this impact estimate is no longer valid for the entire population. Instead, the estimate should be interpreted as a *local* estimate that applies only to a specific subgroup within our target population, the *Enroll-if-assigned*.

Randomized assignment of a program has two important characteristics that allow us to estimate impact when there is imperfect compliance (see box 5.2):

1. It can serve as a predictor of actual enrollment in the program if most people behave as *Enroll-if-assigned*, enrolling in the program when assigned treatment and not enrolling when not assigned treatment.

2. Since the two groups (assigned and not assigned treatment) are generated through a randomized process, the characteristics of individuals in the two groups are not correlated with anything else—such as ability or motivation—that may also affect the outcomes (Y).

Box 5.2: Using Instrumental Variables to Deal with Noncompliance in a School Voucher Program in Colombia

The Program for Extending the Coverage of Secondary School (Programa de Ampliación de Cobertura de la Educación Secundaria, or PACES), in Colombia, provided more than 125,000 students with vouchers covering slightly more than half the cost of attending private secondary school. Because of the limited PACES budget, the vouchers were allocated via a lottery. Angrist and others (2002) took advantage of this randomly assigned treatment to determine the effect of the voucher program on educational and social outcomes.

Angrist and others (2002) found that lottery winners were 10 percent more likely to complete the 8th grade and scored, on average, 0.2 standard deviations higher on standardized tests three years after the initial lottery. They also found that the educational effects were greater for girls than boys. The researchers then looked at the impact of the program on several noneducational outcomes and found that lottery winners were less likely to be married and worked about 1.2 fewer hours per week.

There was some noncompliance with the randomized assignment. Only about 90 percent of the lottery winners actually used the voucher or another form of scholarship, and 24 percent of the lottery losers actually received scholarships. Using our earlier terminology, the population must have contained 10 percent *Never*, 24 percent *Always*, and 66 percent *Enroll-if-assigned*. Angrist and others (2002) therefore also used the original assignment, or a student's lottery win or loss status, as an instrumental variable for the treatment-on-the-treated, or actual receipt of a scholarship. Finally, the researchers were able to calculate a cost-benefit analysis to better understand the impact of the voucher program on both household and government expenditures. They concluded that the total social costs of the program are small and are outweighed by the expected returns to participants and their families, thus suggesting that demand-side programs such as PACES can be a cost-effective way to increase educational attainment.

Source: Angrist and others 2002.

In statistical terms, the randomized assignment serves as an IV. It is a variable that predicts actual enrollment of units in a program, but is not correlated with other characteristics of the units that may be related to outcomes. While some part of the decision of individuals to enroll in a program cannot be controlled by the program administrators, another part of the decision is under their control. In particular, the part of the decision that can be controlled is the assignment to the treatment and comparison groups. Insofar as assignment to the treatment and comparison groups predicts final enrollment in the program, the randomized assignment can be used as an instrument to predict final enrollment. Having this IV allows us to recover the estimates of the local average treatment effect from the estimates of the intention-to-treat effect for the *Enroll-if-assigned* type of units.

A valid IV must satisfy two basic conditions:

1. The IV should not be correlated with the characteristics of the treatment and comparison groups. This is achieved by randomly assigning treatment among the units in the evaluation sample. This is known as *exogeneity*. It is important that the IV not directly affect the outcome of interest. Impacts must be caused only through the program we are interested in evaluating.
2. The IV must affect participation rates in the treatment and comparison groups differently. We typically think of increasing participation in the treatment group. This can be verified by checking that participation is higher in the treatment group compared with the comparison group. This condition is known as *relevance*.

Interpreting the Estimate of the Local Average Treatment Effect

The difference between an estimate of an ATE and an estimate of a LATE is especially important when it comes to interpreting the results of an evaluation. Let's think systematically about how to interpret a LATE estimate. First, we must recognize that individuals who comply in a program (the *Enroll-if-assigned* type) are different from individuals who do not comply (the *Never* and *Always* types). In particular, in the treatment group, noncompliers/nonparticipants (*Never*) may be those who expect to gain little from the intervention. In the comparison group, the noncompliers/participants (*Always*) are likely the group of individuals who expect to benefit the most from participation. In our teacher-training example, teachers who are assigned to the training but decide not to participate (the *Never* type) may be those who feel they don't need training, teachers with a higher opportunity cost of time (for example, because they hold a second job or have children to care for), or teachers with lax supervision who can get away with not attending. On the other hand, teachers who are assigned to the comparison group but enroll anyways (the *Always* type) may be those who feel they absolutely need training, teachers who don't have children of their own to care for, or teachers with a strict principal who insists everyone needs to be trained.

Second, we know that the LATE estimate provides the impact for a particular subgroup of the population: it takes into account only those subgroups that are not affected by either type of noncompliance. In other words, it takes into account only the *Enroll-if-assigned* type. Since the *Enroll-if-assigned* type is different from *Never* and *Always* types, the impact we find through the LATE estimate does not apply to the *Never* or *Always* types. For example, if the ministry of education were to implement a second round of training and somehow force the *Never* teachers who did not get

trained in the first round to get trained, we don't know if those teachers would have lower, equal, or higher effects compared with the teachers who participated in the first round. Similarly, if the most self-motivated teachers always find a way to take the teacher-training program despite being randomly assigned to the comparison group, then the local average treatment effect for the compliers in both treatment and comparison groups does not give us information about the impact of the program for the highly motivated teachers (the *Always*). The estimate of the local average treatment effect applies only to a specific subset of the population: those types that are not affected by noncompliance—that is, only the complier type—and should not be extrapolated to other subsets of the population.

Randomized Promotion as an Instrumental Variable

In the previous section, we saw how to estimate impact based on randomized assignment of treatment, even if compliance with the originally assigned treatment and comparison groups is imperfect. Next we propose a very similar approach that can be applied to evaluate programs that have universal eligibility or open enrollment or in which the program administrator can otherwise not control who participates and who does not.

This approach, called *randomized promotion* (also known as *encouragement design*), provides an additional encouragement for a random set of units to enroll in the program. This randomized promotion serves as an IV. It serves as an external source of variation that affects the probability of receiving the treatment but is otherwise unrelated to the participants' characteristics.

Voluntary enrollment programs typically allow individuals who are interested in the program to decide on their own to enroll and participate. Again consider the job-training program discussed earlier—but this time, randomized assignment is not possible, and any individual who wishes to enroll in the program is free to do so. Very much in line with our previous example, we will expect to encounter different types of people: compliers, a *Never* group, and an *Always* group.

- *Always.* These are the individuals who will always enroll in the program.
- *Never.* These are the individuals who will never enroll.
- *Compliers or Enroll-if-promoted.* In this context, any individual who would like to enroll in the program is free to do so. Yet some individuals may be interested in enrolling but for a variety of reasons, may not have sufficient information or the right incentive to enroll. The compliers here

are those who *enroll-if-promoted*: they are a group of individuals who enroll in the program only if given an additional incentive, stimulus, or promotion that motivates them to enroll. Without this additional stimulus, the *Enroll-if-promoted* would simply remain out of the program.

Returning to the job-training example, if the agency that organizes the training is well funded and has sufficient capacity, it may have an "open-door" policy, treating every unemployed person who wants to participate. It is unlikely, however, that every unemployed person will actually step forward to participate or will even know that the program exists. Some unemployed people may be reluctant to enroll because they know very little about the content of the training and find it hard to obtain additional information. Now assume that the job-training agency hires a community outreach worker to go around town to encourage a randomly selected group of unemployed persons to enroll into the job-training program. Carrying the list of randomly selected unemployed people, she knocks on their doors, describes the training program, and offers to help the person to enroll in the program on the spot. The visit is a form of promotion, or encouragement, to participate in the program. Of course, she cannot force anyone to participate. In addition, the unemployed persons whom the outreach worker does not visit can also enroll, although they will have to go to the agency themselves to do so. So we now have two groups of unemployed people: those who were randomly assigned a visit by the outreach worker, and those who were randomly not visited. If the outreach effort is effective, the enrollment rate among unemployed people who were visited should be higher than the rate among unemployed people who were not visited.

Now let us think about how we can evaluate this job-training program. We cannot just compare those unemployed people who enroll with those who do not enroll. That's because the unemployed who enroll are probably very different from those who do not enroll in both observed and unobserved ways: they may be more or less educated (this can be observed easily), and they are probably more motivated and eager to find a job (this is hard to observe and measure).

However, there is some additional variation that we can exploit to find a valid comparison group. Consider for a moment whether we can compare the group of people who were randomly assigned to receive a visit from the outreach worker with the group that was not visited. Because the promoted and nonpromoted groups were determined at random, both groups contain identical compositions of very motivated persons *(Always)* who will enroll whether or not the outreach worker knocks on their door. Both groups also contain unmotivated persons *(Never)* who will not enroll in the program, despite the efforts of the outreach worker. Finally, if the outreach worker is

effective at motivating enrollment, some people (*Enroll-if-promoted*) will enroll in the training if the outreach worker visits them, but will not enroll if the worker does not.

Since the outreach worker visited a group of individuals assigned at random, we can derive a LATE estimate, as discussed earlier. The only difference is that instead of randomly *assigning* the program, we are randomly *promoting* it. As long as *Enroll-if-promoted* people (who enroll when we reach out to them but do not enroll when we do not reach out to them) appear in sufficient numbers, we have variation between the group *with* the promotion and the group *without* the promotion that allows us to identify the impact of the training on the *Enroll-if-promoted*. Instead of complying with the assignment of the treatment, the *Enroll-if-promoted* are now complying with the promotion.

For this strategy to work, we want the outreach or promotion to be effective in increasing enrollment substantially among the *Enroll-if-promoted* group. At the same time, we do not want the promotion activities themselves to influence the final outcomes of interest (such as earnings), since at the end of the day we are interested primarily in estimating the impact of the training program, and not the impact of the promotion strategy, on final outcomes. For example, if the outreach workers offered large amounts of money to unemployed people to get them to enroll, it would be hard to tell whether any later changes in income were caused by the training or by the outreach activity itself.

Randomized promotion is a creative strategy that generates the equivalent of a comparison group for the purposes of impact evaluation. It can be used when a program has open enrollment and it is feasible to organize a promotion campaign aimed at a random sample of the population of interest. Randomized promotion is another example of an IV that allows us to estimate impact in an unbiased way. But again, as with randomized assignment with imperfect compliance, impact evaluations relying on randomized promotion provide a LATE estimate: a local estimate of the effect on a specific subgroup of the population, the *Enroll-if-promoted* group. As before, this LATE estimate cannot be directly extrapolated to the whole population, since the *Always* and *Never* groups are likely quite different from the *Enroll-if-promoted* group.

Key Concept

Randomized promotion is an instrumental variable method that allows us to estimate impact in an unbiased way. It randomly assigns a promotion, or encouragement, to participate in the program. It is a useful strategy to evaluate programs that are open to everyone who is eligible.

You Said "Promotion"?

Randomized promotion seeks to increase the take-up of a voluntary program in a randomly selected subsample of the population. The promotion itself can take several forms. For instance, we may choose to initiate an information campaign to reach those individuals who had not

enrolled because they did not know or fully understand the content of the program. Alternatively, we may choose to provide incentives to sign up, such as offering small gifts or prizes or making transportation available.

As discussed for IV more generally, a number of conditions must be met for the randomized promotion approach to produce valid estimate of program impact:

1. The promoted and nonpromoted groups must be similar. That is, the average characteristics of the two groups must be statistically equivalent. This is achieved by randomly assigning the outreach or promotion activities among the units in the evaluation sample.
2. The promotion itself should not directly affect the outcomes of interest. This is a critical requirement so that we can tell that changes in the outcomes of interest are caused by the program itself and not by the promotion.
3. The promotion campaign must substantially change enrollment rates in the promoted group relative to the nonpromoted group. We typically think of increasing enrollment with promotion. This can be verified by checking that enrollment rates are higher in the group that receives the promotion than in the group that does not.

The Randomized Promotion Process

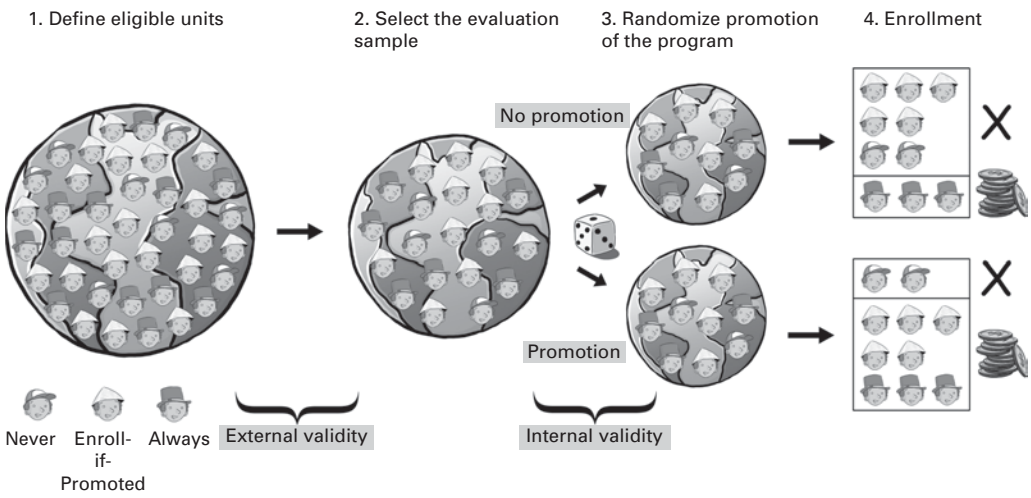
The process of randomized promotion is presented in figure 5.3. As in the previous methods, we begin with the population of eligible units for the program. In contrast with randomized assignment, we can no longer randomly choose who will receive the program and who will not receive the program because the program is fully voluntary. However, within the population of eligible units, there will be three types of units:

- *Always.* Those who will always want to enroll in the program.
- *Enroll-if-promoted.* Those who will sign up for the program only when given additional promotion.
- *Never.* Those who never want to sign up for the program, whether or not we offer them promotion.

Again, note that being an *Always,* an *Enroll-if-promoted,* or a *Never* is an intrinsic characteristic of units that cannot be easily measured by the program evaluation team because it is related to factors such as motivation, intelligence, and information.

Once the eligible population is defined, the next step is to randomly select a sample from the population to be part of the evaluation. These are

Figure 5.3 Randomized Promotion



the units on whom we will collect data. In some cases—for example, when we have data for the entire population of eligible units—we may decide to include this entire population in the evaluation sample.

Once the evaluation sample is defined, randomized promotion randomly assigns the evaluation sample into a promoted group and a nonpromoted group. Since we are randomly choosing the members of both the promoted group and the nonpromoted group, both groups will share the characteristics of the overall evaluation sample, and those will be equivalent to the characteristics of the population of eligible units. Therefore, the promoted group and the nonpromoted group will have similar characteristics.

After the promotion campaign is over, we can observe the enrollment rates in both groups. In the nonpromoted group, only the *Always* will enroll. Although we know which units are *Always* in the nonpromoted group, we will not be able to distinguish between the *Never* and *Enroll-if-promoted* in that group. By contrast, in the promoted group, both the *Enroll-if-promoted* and the *Always* will enroll, whereas the *Never* will not enroll. So in the promoted group we will be able to identify the *Never* group, but we will not be able to distinguish between the *Enroll-if-promoted* and the *Always*.

Estimating Impact under Randomized Promotion

Imagine that for a group of 10 individuals per group, the promotion campaign raises enrollment from 30 percent in the nonpromoted group (3 *Always*) to 80 percent in the promoted group (3 *Always* and 5 *Enroll-if-promoted*). Assume that the average outcome for all individuals the

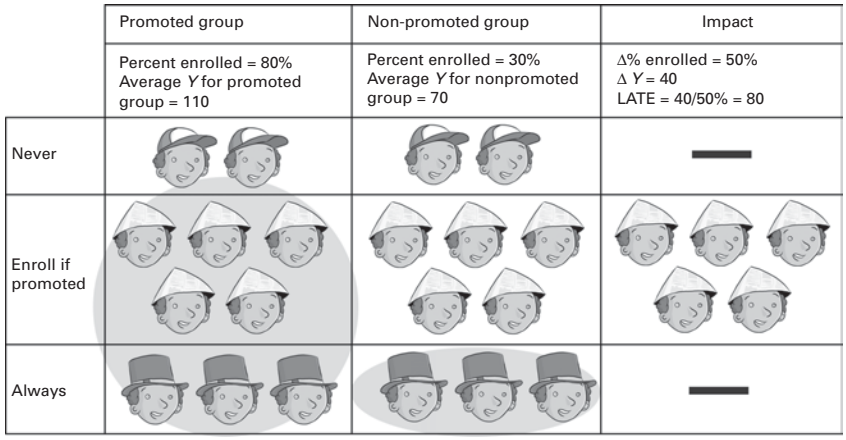
nonpromoted group (10 individuals) is 70, and that average outcome for all individuals in the promoted group (10 individuals) is 110 (figure 5.4). Then what would be the impact of the program?

First, let's compute the straight difference in outcomes between the promoted and the nonpromoted groups, which is 40 (110 minus 70). We know that none of this difference of 40 comes from the *Nevers* because they do not enroll in either group. We also know that none of this difference of 40 should come from the *Always* because they enroll in both groups. So all of this difference of 40 should come from the *Enroll-if-promoted*.

The second step is to obtain the LATE estimate of the program on the *Enroll-if-promoted*. We know that the entire difference between the promoted and nonpromoted groups of 40 can be attributed to the *Enroll-if-promoted*, who make up only 50 percent of the population. To assess the average effect of the program on a complier, we divide 40 by the percentage of *Enroll-if-promoted* in the population. Although we cannot directly identify the *Enroll-if-promoted*, we are able to deduce what must be their *percentage* of the population: it is the difference in the enrollment rates of the promoted and the nonpromoted groups (50 percent, or 0.5). Therefore, the estimate of the local average treatment effect of the program on the *Enroll-if-promoted* group is 40/0.5=80.

Given that the promotion is assigned randomly, the promoted and nonpromoted groups have equal characteristics. Thus the differences that we observe in average outcomes between the two groups must be caused by

Figure 5.4 Estimating the Local Average Treatment Effect under Randomized Promotion



Note: Δ = causal impact; Y = outcome. Characters that appear against the shaded background are those who enroll.

the fact that in the promoted group, the *Enroll-if-promoted* enroll, while in the nonpromoted group, they do not. Again, we should not directly extrapolate the estimated impacts for the *Enroll-if-promoted* to other groups, since they are likely quite different from the groups that *Never* and *Always* enroll. Box 5.3 presents an example of randomized promotion for a project in Bolivia.

Box 5.3: Randomized Promotion of Education Infrastructure Investments in Bolivia

In 1991, Bolivia institutionalized and scaled up a successful Social Investment Fund (SIF), which provided financing to rural communities to carry out small-scale investments in education, health, and water infrastructure. The World Bank, which was helping to finance SIF, built an impact evaluation into the program design.

As part of the impact evaluation of the education component, communities in the Chaco region were randomly selected for active promotion of the SIF intervention and received additional visits and encouragement to apply from program staff. The program was open to all eligible communities in the region and was demand-driven, in that communities had to apply for funds for a specific project. Not all communities took up the program, but take-up was higher among promoted communities.

Newman and others (2002) used the randomized promotion as an instrumental variable. They found that the education investments succeeded in improving measures of school infrastructure quality such as electricity, sanitation facilities, textbooks per student, and student-teacher ratios. However, they detected little impact on educational outcomes, except for a decrease of about 2.5 percent in the dropout rate. As a result of these findings, the ministry of education and the SIF now focus more attention and resources on the "software" of education, funding physical infrastructure improvements only when they form part of an integrated intervention.

Source: Newman and others 2002.

Evaluating the Impact of HISP: Randomized Promotion

Let us now try using the randomized promotion method to evaluate the impact of the Health Insurance Subsidy Program (HISP). Assume that the ministry of health makes an executive decision that the health insurance subsidy should be made available immediately to any household that wants to enroll. You note that this is a different scenario than the randomized assignment case we have considered so far. However, you know that realistically this national scale-up will be incremental over

time, so you reach an agreement to try and accelerate enrollment in a random subset of villages through a promotion campaign. In a random subsample of villages, you undertake an intensive promotion effort that includes communication and social marketing aimed at increasing awareness of HISP. The promotion activities are carefully designed to avoid content that may inadvertently encourage changes in other health-related behaviors, since this would invalidate the promotion as an instrumental variable (IV). Instead, the promotion concentrates exclusively on boosting enrollment in HISP. After two years of promotion and program implementation, you find that 49.2 percent of households in villages that were randomly assigned to the promotion have enrolled in the program, while only 8.4 percent of households in nonpromoted villages have enrolled (table 5.1).

Because the promoted and nonpromoted villages were assigned at random, you know that the average characteristics of the two groups should be the same in the absence of the promotion. You can verify that assumption by comparing the baseline health expenditures (as well as any other characteristics) of the two populations. After two years of program implementation, you observe that the average health expenditure in the promoted villages is US$14.97, compared with US$18.85 in nonpromoted areas (a difference of minus US$3.87). However, because the only difference between the promoted and nonpromoted villages is that enrollment in the program is higher in the promoted villages (thanks to the promotion), this difference of US$3.87 in health expenditures must be due to the additional 40.78 percent of households that enrolled in the promoted villages because of the promotion. Therefore, we need to adjust the difference in health expenditures to be able to find the impact of the program on the *Enroll-if-promoted*. To do this, we divide the intention-to-treat estimate—that is, the straight difference between the promoted and nonpromoted groups—by the percentage of *Enroll-if-promoted:* −3.87/0.4078 = −US$9.49.

Table 5.1 Evaluating HISP: Randomized Promotion Comparison of Means

	Promoted villages	Nonpromoted villages	Difference	*t*-stat
Household health expenditures at baseline (US$)	17.19	17.24	−0.05	−0.47
Household health expenditures at follow-up (US$)	14.97	18.85	−3.87	−16.43
Enrollment rate in HISP	49.20%	8.42%	40.78%	49.85

Note: Significance level: ** = 1 percent.

Table 5.2 Evaluating HISP: Randomized Promotion with Regression Analysis

	Linear regression	Multivariate linear regression
Estimated impact on household health expenditures (US$)	–9.50** (0.52)	–9.74** (0.46)

Note: Standard errors are in parentheses. Significance level: ** = 1 percent.

Your colleague, an econometrician who suggests using the randomized promotion as an IV, then estimates the impact of the program through a two-stage least-squares procedure (see online technical companion at http://www.worldbank.org/ieinpractice for further details on the econometric approach to estimating impacts with IV). She finds the results shown in table 5.2. This estimated impact is valid for those households that enrolled in the program because of the promotion but who otherwise would not have done so: in other words, the *Enroll-if-promoted*.

HISP Question 4

A. What are the key conditions required to accept the results from the randomized promotion evaluation of HISP?
B. Based on these results, should HISP be scaled up nationally?

Limitations of the Randomized Promotion Method

Randomized promotion is a useful strategy for evaluating the impact of voluntary programs and programs with universal eligibility, particularly because it does not require the exclusion of any eligible units. Nevertheless, the approach has some noteworthy limitations compared with randomized assignment of treatment.

First, the promotion strategy must be effective. If the promotion campaign does not increase enrollment, then no difference between the promoted and the nonpromoted groups will appear, and there will be nothing to compare. It is thus crucial to carefully design and extensively pilot the promotion campaign to make sure that it will be effective. On the positive side, the design of the promotion campaign can help program managers by teaching them how to increase enrollment after the evaluation period is concluded.

Second, the randomized promotion method estimates the impact of the program for only a subset of the population of eligible units (a LATE).

Specifically, the program's local average impact is estimated from the group of individuals who sign up for the program only when encouraged to do so. However, individuals in this group may have very different characteristics than those individuals who always or never enroll. Therefore the average treatment effect for the entire population may be different from the average treatment effect estimated for individuals who participate only when encouraged. A randomized promotion evaluation will not estimate impacts for the group of individuals who enroll in the program without encouragement. In some contexts, this group (the *Always*) may be precisely the group the program is designed to benefit. In this context, the randomized promotion design will shed light on impacts expected for new populations that would enroll from additional promotion, but not on impacts for the population that already enrolls on its own.

Checklist: Randomized Promotion as an Instrumental Variable

Randomized promotion leads to valid estimates of the counterfactual if the promotion campaign substantially increases take-up of the program without directly affecting the outcomes of interest.

- ✓ Are the baseline characteristics balanced between the units that received the promotion campaign and those that did not? Compare the baseline characteristics of the two groups.
- ✓ Does the promotion campaign substantially affect the take-up of the program? It should. Compare the program take-up rates in the promoted and the nonpromoted subsamples.
- ✓ Does the promotion campaign directly affect outcomes? It should not. This cannot usually be directly tested, so you need to rely on theory, common sense, and good knowledge of the setting of the impact evaluation for guidance.

Additional Resources

- For accompanying material to the book and hyperlinks to additional resources, please see the Impact Evaluation in Practice website (http://www.worldbank.org/ieinpractice).
- For additional resources on IV, see the Inter-American Development Bank Evaluation Portal (http://www.iadb.org/evaluationhub).

Notes

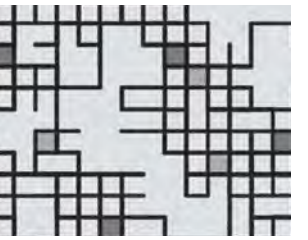
1. In the medical sciences, patients in the comparison group typically receive a placebo: that is, something like a sugar pill that should have no effect on the intended outcome. That is done to further control for the *placebo effect*, meaning the potential changes in behavior and outcomes that could occur simply from the act of receiving a treatment, even if the treatment itself is ineffective.
2. These two steps correspond to the econometric technique of two-stage least-squares, which produces an estimate of the local average treatment effect.

References

Angrist, Joshua, Eric Bettinger, Erik Bloom, Elizabeth King, and Michael Kremer. 2002. "Vouchers for Private Schooling in Colombia: Evidence from a Randomized Natural Experiment." *American Economic Review* 92 (5): 1535–58.

Kearney, Melissa S., and Philip B. Levine. 2015. "Early Childhood Education by MOOC: Lessons from *Sesame Street*." NBER Working Paper 21229, National Bureau of Economic Research, Cambridge, MA.

Newman, John, Menno Pradhan, Laura B. Rawlings, Geert Ridder, Ramiro Coa, and Jose Luis Evia. 2002. "An Impact Evaluation of Education, Health, and Water Supply Investments by the Bolivian Social Investment Fund." *World Bank Economic Review* 16 (2): 241–74.



CHAPTER 6

Regression Discontinuity Design

Evaluating Programs That Use an Eligibility Index

Social programs often use an index to decide who is eligible to enroll in the program and who is not. For example, antipoverty programs are typically targeted to poor households, which are identified by a poverty score or index. The poverty score can be based on a formula that measures a set of basic household assets as a proxy (or estimate) for means (such as income, consumption, or purchasing power).[1] Households with low scores are classified as poor, and households with higher scores are considered relatively better-off. Antipoverty programs typically determine a threshold or cutoff score, below which households are deemed poor and are eligible for the program. Colombia's system for selecting beneficiaries of social spending is one such example (see box 6.1). Test scores are another example (see box 6.3). College admission might be granted to the top performers on a standardized test, whose results are ranked from the lowest to the highest performer. If the number of slots is limited, then only students who score above a certain threshold score (such as the top 10 percent of students) will be granted admission. In both examples, there is a continuous eligibility index (poverty score and test score, respectively) that allows for ranking the population of interest, as well as a threshold or cutoff score that determines who is eligible and who is not.

Box 6.1: Using Regression Discontinuity Design to Evaluate the Impact of Reducing School Fees on School Enrollment Rates in Colombia

Barrera-Osorio, Linden, and Urquiola (2007) used regression discontinuity design (RDD) to evaluate the impact of a school fee reduction program in Colombia (Gratuitad) on school enrollment rates in the city of Bogota. The program is targeted based on an index called the SISBEN, which is a continuous poverty index whose value is determined by household characteristics, such as location, the building materials of the home, the services that are available there, demographics, health, education, income, and the occupations of household members. The government established two cutoff scores along the SISBEN index: children of households with scores below cutoff score no. 1 are eligible for free education from grades 1 to 11; children of households with scores between cutoff scores no. 1 and no. 2 are eligible for a 50 percent subsidy on fees for grades 10 and 11; and children from households with scores above cutoff score no. 2 are not eligible for free education or subsidies.

The authors used a RDD for four reasons. First, household characteristics such as income or the education level of the household head are continuous along the SISBEN score at baseline; in other words, there are no "jumps" in characteristics along the SISBEN score. Second, households on both sides of the cutoff scores have similar characteristics, generating credible comparison groups. Third, a large sample of households was available. Finally, the government kept the formula used to calculate the SISBEN index secret, so that scores would be protected from manipulation.

Using the RDD method, the researchers found that the program had a significant positive impact on school enrollment rates. Specifically, enrollment was 3 percentage points higher for primary school students from households below cutoff score no. 1, and 6 percentage points higher for high school students from households between cutoff scores no. 1 and no. 2. This study provides evidence on the benefits of reducing the direct costs of schooling, particularly for at-risk students. However, its authors also call for further research on price elasticities to better inform the design of subsidy programs such as this one.

Source: Barrera-Osorio, Linden, and Urquiola 2007.

Regression discontinuity design (RDD) is an impact evaluation method that can be used for programs that have a continuous eligibility index with a clearly defined eligibility threshold (cutoff score) to determine who is eligible and who is not. To apply a regression discontinuity design, the following main conditions must be met:

1. The index must rank people or units in a continuous or "smooth" way. Indexes like poverty scores, test scores, or age have many values that can be ordered from small to large, and therefore they can be considered smooth. By contrast, variables that have discrete or "bucket" categories

that have only a few possible values or cannot be ranked are not considered smooth. Examples of the latter include employment status (employed or unemployed), highest education level achieved (primary, secondary, university, or postgraduate), car ownership (yes or no), or country of birth.

2. The index must have a clearly defined cutoff score: that is, a point on the index above or below which the population is classified as eligible for the program. For example, households with a poverty index score of less than 50 out of 100 might be classified as poor, individuals age 67 and older might be classified as eligible for a pension, and students with a test score of 90 or more out of 100 might be eligible for a scholarship. The cutoff scores in these examples are 50, 67, and 90, respectively.
3. The cutoff must be unique to the program of interest; that is, there should be no other programs, apart from the program to be evaluated, that uses the same cutoff score. For example, if a poverty score below 50 qualifies a household for a cash transfer, health insurance, and free public transportation, we would not be able to use the RDD method to estimate the impact of the cash transfer program by itself.
4. The score of a particular individual or unit cannot be manipulated by enumerators, potential beneficiaries, program administrators, or politicians.

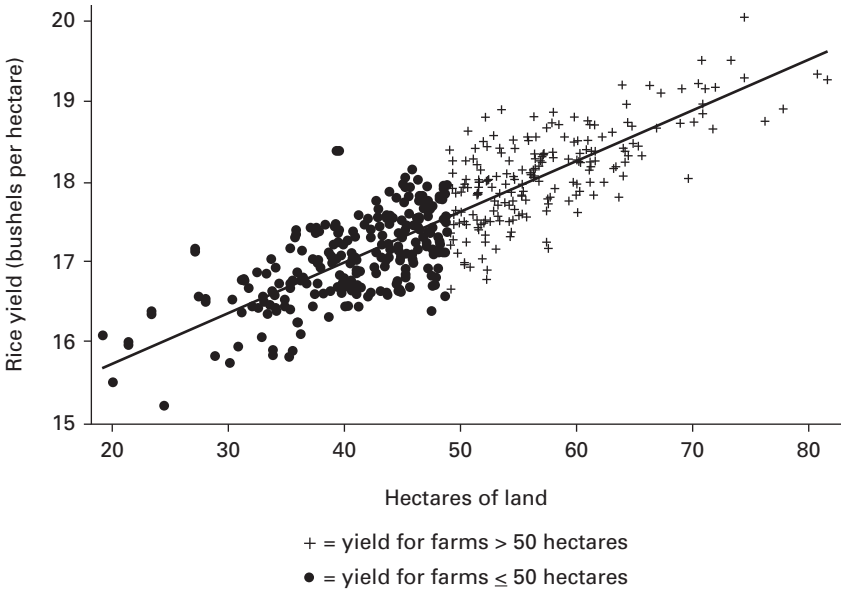
The RDD estimates impact around the eligibility cutoff as the difference between the average outcome for units on the treated side of the eligibility cutoff and the average outcome of units on the untreated (comparison) side of the cutoff.

Key Concept

Regression discontinuity design (RDD) is an impact evaluation method that is adequate for programs that use a continuous index to rank potential participants and that have a cutoff point along the index that determines whether or not potential participants are eligible to receive the program.

Consider an agriculture program that aims to improve total rice yields by subsidizing farmers' purchase of fertilizer. The program targets small and medium-size farms, which it classifies as farms with fewer than 50 hectares of land. Before the program starts, we might expect smaller farms to have lower outputs than larger farms, as shown in figure 6.1, which plots farm size and rice production. The eligibility score in this case is the number of hectares of the farm, and the cutoff is 50 hectares. Program rules establish that farms below the 50-hectare cutoff are eligible to receive fertilizer subsidies, and farms with 50 or more hectares are not. In this case, we might expect to see a number of farms with 48, 49, or even 49.9 hectares that participate in the program. Another group of farms with 50, 50.1, and 50.2 hectares will not participate in the program because they lie just to the ineligible side of the cutoff. The group of farms with 49.9 hectares is likely to be very similar to the group of farms with 50.1 hectares in all respects, except that one group received the fertilizer subsidy and the other group did not. As we move further away from the eligibility cutoff, eligible and

Figure 6.1 Rice Yield, Smaller Farms versus Larger Farms (Baseline)

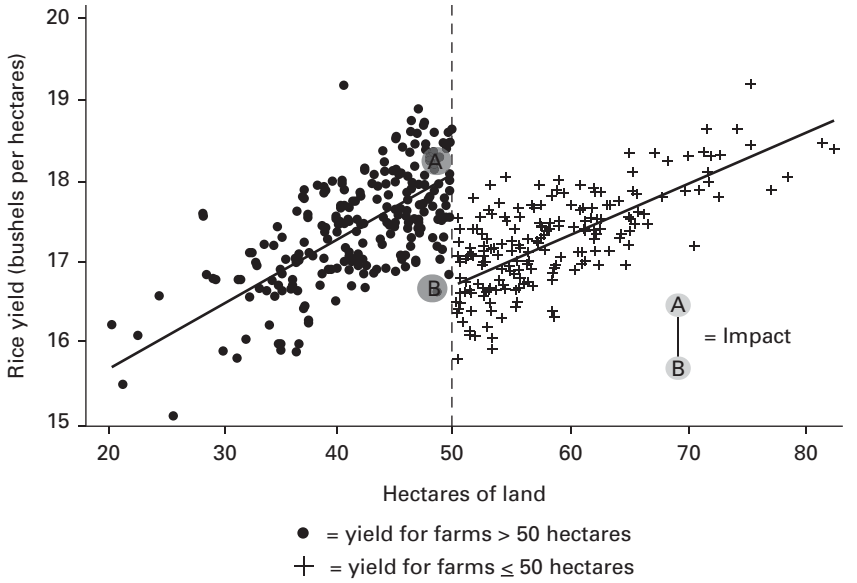


ineligible farms may differ more. But farm size is a good measure of how different they are, allowing us to control for many of those differences.

Once the program rolls out and subsidizes the cost of fertilizer for small and medium farms, the impact evaluation could use an RDD to evaluate its impact (figure 6.2). The RDD calculates impact as the difference in outcomes, such as rice yields, between the units on both sides of the eligibility cutoff, which in our example is a farm size of 50 hectares. The farms that were just too large to enroll in the program constitute the comparison group and generate an estimate of the counterfactual outcome for those farms in the treatment group that were just small enough to enroll. Given that these two groups of farms were very similar at baseline and are exposed to the same set of external factors over time (such as weather, price shocks, and local and national agricultural policies), the only plausible reason for different outcomes must be the program itself.

Since the comparison group is made up of farms just above the eligibility threshold, the impact given by a RDD is valid only locally—that is, in the neighborhood around the eligibility cutoff score. Thus we obtain an estimate of a local average treatment effect (LATE) (see chapter 5). The impact of the fertilizer subsidy program is valid for the larger of the medium-size farms: that is, those with just under 50 hectares of land. The impact evaluation will not necessarily be able to directly identify the impact of the

Figure 6.2 Rice Yield, Smaller Farms versus Larger Farms (Follow-Up)



program on the smallest farms—say, those with 10 or 20 acres of land—where the effects of a fertilizer subsidy may differ in important ways from the medium-size farms with 48 or 49 hectares. One advantage of the RDD method is that once the program eligibility rules are applied, no eligible units need to be left untreated for the purposes of the impact evaluation. The trade-off is that impacts for observations far away from the cutoff will not be known. Box 6.2 presents an example of the use of RDD for evaluating a social safety net program in Jamaica.

Fuzzy Regression Discontinuity Design

Once we have verified that there is no evidence of manipulation in the eligibility index, we may still face a challenge if units do not respect their assignment to the treatment or comparison groups. In other words, some units that qualify for the program on the basis of their eligibility index may opt not to participate, while other units that did not qualify for the program on the basis of their eligibility index may find a way to participate anyway. When all units comply with the assignment that corresponds to them on the basis of their eligibility index, we say that the RDD is "sharp," while if there is noncompliance on either side of the cutoff, then

Box 6.2: Social Safety Nets Based on a Poverty Index in Jamaica

The regression discontinuity design (RDD) method was used to evaluate the impact of a social safety net initiative in Jamaica. In 2001, the government of Jamaica initiated the Programme of Advancement through Health and Education (PATH) to increase investments in human capital and improve the targeting of welfare benefits to the poor. The program provided health and education grants to children in eligible poor households, conditional on school attendance and regular health care visits. The average monthly benefit for each child was about US$6.50, in addition to a government waiver of certain health and education fees.

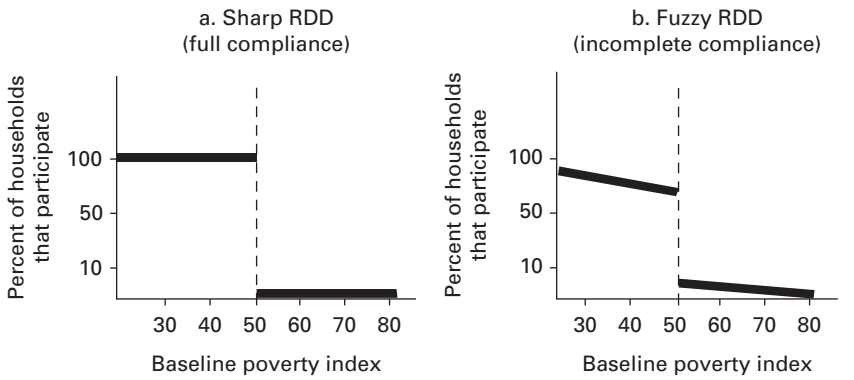
With program eligibility determined by a scoring formula, Levy and Ohls (2010) were able to compare households just below the eligibility threshold with households just above (between 2 and 15 points from the cutoff). The researchers justify using the RDD method with baseline data showing that the treatment and comparison households had similar levels of poverty, measured by proxy means scores, and similar levels of motivation, in that all of the households in the sample had applied to the program. The researchers also used the program eligibility score in the regression analysis to help control for any differences between the two groups.

Levy and Ohls (2010) found that the PATH program increased school attendance for children ages 6 to 17 by an average of 0.5 days per month, which is significant given an already fairly high attendance rate of 85 percent. Moreover, health care visits by children ages 0 to 6 increased by approximately 38 percent. While the researchers were unable to find any longer-term impacts on school achievement or health care status, they concluded that the magnitude of the impacts they did find was broadly consistent with conditional cash transfer programs implemented in other countries. A final interesting aspect of this evaluation is that it gathered both quantitative and qualitative data, using information systems, interviews, focus groups, and household surveys.

Source: Levy and Ohls 2010.

we say that the RDD is "fuzzy" (figure 6.3). If the RDD is fuzzy, we can use the instrumental variable approach to correct for the noncompliance (see chapter 5). Remember that in the case of randomized assignment with noncompliance, we used the randomized assignment as the instrumental variable that helped us correct for noncompliance. In the case of RDD, we can use the original assignment based on the eligibility index as the instrumental variable. Doing so has a drawback, though: our instrumental RDD impact estimate will be further localized—in the sense that it is no longer valid to all observations close to the cutoff, but instead represents the impact for the subgroup of the population that is located close to the cutoff point and that participates in the program only because of the eligibility criteria.

Figure 6.3 Compliance with Assignment

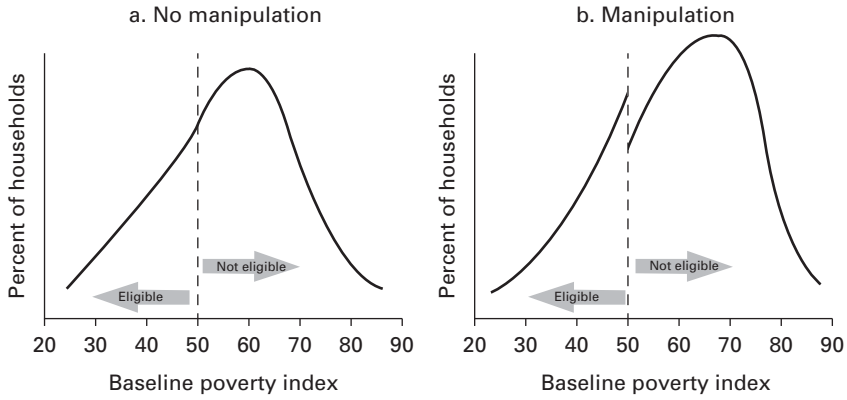


Checking the Validity of the Regression Discontinuity Design

For a RDD to yield an unbiased LATE estimate at the cutoff, it is important that the eligibility index not be manipulated around the cutoff so that an individual can change treatment or control status.[2] Manipulation of the eligibility criteria can take many forms. For example, the enumerators who collect data that are used to compute the eligibility score could change one or two responses of respondents; or respondents may purposefully lie to enumerators if they think that doing so would qualify them for the program. In addition, manipulation of the scores might get worse over time as enumerators, respondents, and politicians all start learning the "rules of the game." In the fertilizer subsidy example, manipulation around the cutoff would occur if farm owners could alter land titles or misreport the size of their farms. Or a farmer with 50.3 hectares of land might find a way to sell off a half hectare to qualify for the program, if the expected benefits from the fertilizer subsidy were worth doing so.

One telltale sign of manipulation is illustrated in figure 6.4. Panel a shows the distribution of households according to their baseline index when there is no manipulation. The density of households around the cutoff (50) is continuous (or smooth). Panel b shows a different situation: a larger number of households seem to be "bunched" right below the cutoff, while relatively few households can be found right above the cutoff. Since there is no a priori reason to believe that there should be a large shift in the number of households right around the cutoff, the occurrence of that shift in the distribution around the cutoff is evidence that somehow households

Figure 6.4 Manipulation of the Eligibility Index



Box 6.3: The Effect on School Performance of Grouping Students by Test Scores in Kenya

To test whether assigning students to classes based on performance improves educational outcomes, Duflo, Dupas, and Kremer (2011) conducted an experiment with 121 primary schools in western Kenya. In half the schools, first-grade students were randomly split into two different class sections. In the other half of the schools, students were assigned to either a high-performing or a low-performing section based on their initial test scores, using the test score as a cutoff point.

The regression discontinuity design (RDD) allowed researchers to test whether the composition of students in a class directly affected test scores. They compared endline test scores for students who were right around the cutoff to see if those assigned to the high-performing section did better than those assigned to the low-performing section.

On average, endline test scores in schools that assigned students to sections with similarly higher or lower performers were 0.14 standard deviations higher than in schools that did not use this method and instead used randomized assignment to create equivalent groups of students. These results were not solely driven by students in the high-performing section, as students in the low-performing section also showed improvements in test scores. For students right around the cutoff score, the researchers found that there was no significant difference in endline test scores. These findings reject the hypothesis that students directly benefit from having higher-achieving classmates.

Source: Duflo, Dupas, and Kremer 2011.

may be manipulating their scores to gain access to the program. A second test for manipulation plots the eligibility index against the outcome variable at baseline and checks that there is no discontinuity or "jump" right around the cutoff line.

Evaluating the Impact of HISP: Regression Discontinuity Design

Now consider how the regression discontinuity design (RDD) method can be applied to our Health Insurance Subsidy Program (HISP). After doing some more investigation into the design of HISP, you find that in addition to randomly selecting treatment villages, the authorities targeted the program to low-income households using the national poverty line. The poverty line is based on a poverty index that assigns each household in the country a score between 20 and 100 based on its assets, housing conditions, and sociodemographic structure. The poverty line has been officially set at 58. This means that all households with a score of 58 or below are classified as poor, and all households with a score of more than 58 are considered to be nonpoor. Even in the treatment villages, only poor households are eligible to enroll in HISP. Your data set includes information on both poor and nonpoor households in the treatment villages.

Before carrying out the regression discontinuity design estimations, you decide to check whether there is any evidence of manipulation of the eligibility index. As a first step, you check whether the density of the eligibility index raises any concerns about manipulation of the index. You plot the percentage of households against the baseline poverty index (figure 6.5).[3] The figure does not indicate any "bunching" of households right below the cutoff of 58.

Next, you check whether households respected their assignment to the treatment and comparison groups on the basis of their eligibility score. You plot participation in the program against the baseline poverty index (figure 6.6) and find that two years after the start of the pilot, only households with a score of 58 or below (that is, to the left of the poverty line) have been allowed to enroll in HISP. In addition, all of the eligible households enrolled in HISP. In other words, you find full compliance and have a "sharp" RDD.

You now proceed to apply the RDD method to compute the impact of the program. Using follow-up data, you again plot the relationship between the scores on the poverty index and predicted health

Figure 6.5 HISP: Density of Households, by Baseline Poverty Index

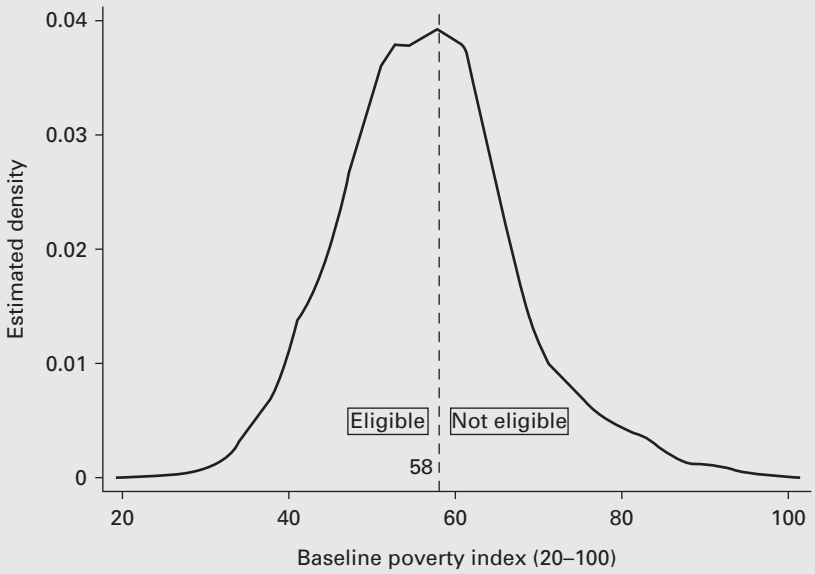


Figure 6.6 Participation in HISP, by Baseline Poverty Index

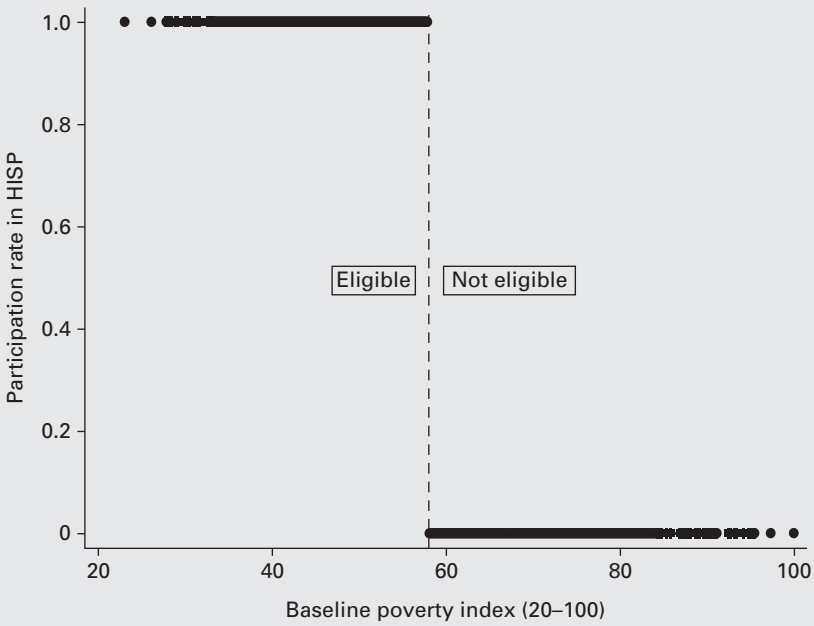


Figure 6.7 Poverty Index and Health Expenditures, HISP, Two Years Later

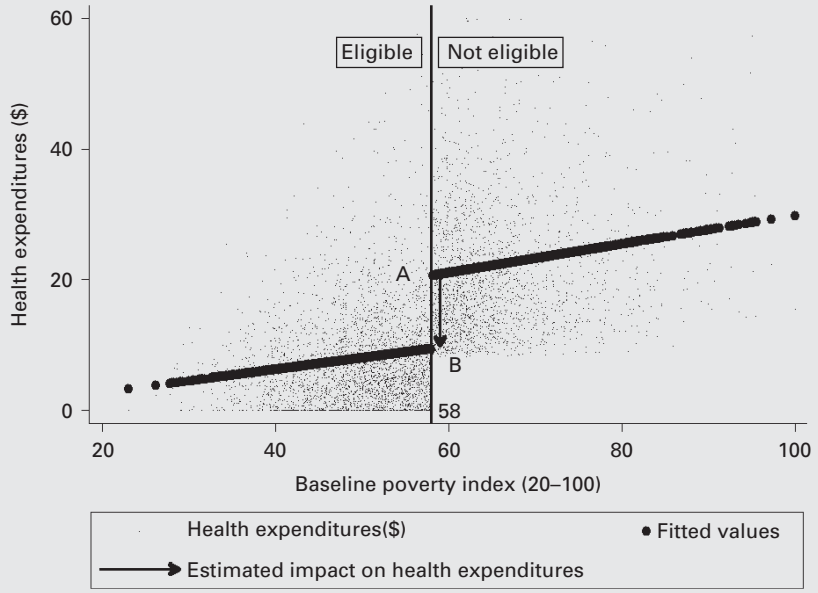


Table 6.1 Evaluating HISP: Regression Discontinuity Design with Regression Analysis

	Multivariate linear regression
Estimated impact on household health expenditures	–9.03**
	(0.43)

Note: Standard errors are in parentheses. Significance level: ** = 1 percent.

expenditures and find the relation illustrated in figure 6.7. In the relationship between the poverty index and the predicted health expenditures, you find a clear break, or *discontinuity*, at the poverty line (58).

The discontinuity reflects a decrease in health expenditures for those households eligible to receive the program. Given that households on both sides of the cutoff score of 58 are very similar, the plausible explanation for the different level of health expenditures is that one group of households was eligible to enroll in the program and the other was not. You estimate this difference through a regression with the findings shown in table 6.1.

HISP Question 5

A. Is the result shown in table 6.1 valid for all eligible households?
B. Compared with the impact estimated with the randomized assignment method, what does this result say about those households with a poverty index of just under 58?
C. Based on the RDD impact estimates, should HISP be scaled up nationally?

Limitations and Interpretation of the Regression Discontinuity Design Method

Regression discontinuity design provides estimates of local average treatment effects (LATE) around the eligibility cutoff at the point where treatment and comparison units are most similar. The closer to the cutoff you get, the more similar the units on either side of the cutoff will be. In fact, when you get extremely close to the cutoff score, the units on either side of the cutoff will be so similar that your comparison will be as good as if you had chosen the treatment and comparison groups using randomized assignment of the treatment.

Because the RDD method estimates the impact of the program around the cutoff score, or *locally*, the estimate cannot necessarily be generalized to units whose scores are further away from the cutoff score: that is, where eligible and ineligible individuals may not be as similar. The fact that the RDD method will not be able to provide an estimate of an average treatment effect for all program participants can be seen as both a strength and a limitation of the method, depending on the evaluation question of interest. If the evaluation primarily seeks to answer the question, should the program exist or not?, then the average treatment effect for the entire eligible population may be the most relevant parameter, and clearly the RDD will fall short of being perfect. However, if the policy question of interest is, should the program be cut or expanded at the margin?—that is, for (potential) beneficiaries right around the cutoff—then the RDD produces precisely the local estimate of interest to inform this important policy decision.

As mentioned, there can be an additional complication when compliance on either side of the cutoff is imperfect. This fuzzy RDD happens when units that are not eligible based on their index score nonetheless manage to gain access to the program, or when units that are eligible based on their index score choose not to participate in the program. In this case,

we can use an instrumental variable methodology that is similar to the one outlined in chapter 5: the location of units above or below the cutoff score will be used as an instrumental variable for the observed participation in the program. As was the case in the examples discussed in chapter 5, doing this has a drawback: we can estimate the impact for only those units that are sensitive to the eligibility criteria—the *Enroll-if-eligible-score* type, not the *Always* or *Never* types.

The fact that the RDD method estimates impact only around the cutoff score also raises challenges in terms of the statistical power of the analysis. Sometimes only a restricted set of observations that are located close to the cutoff score are used in the analysis, thereby lowering the number of observations in the RDD analysis relative to methods that analyze all units in the treatment and comparison groups. To obtain sufficient statistical power when applying RDD, you will need to choose a *bandwidth* around the cutoff score that includes a sufficient number of observations. In practice, you should try to use as large a bandwidth as possible, while maintaining the balance in observed characteristics of the population above and below the cutoff score. You can then run the estimation several times using different bandwidths to check whether the estimates are sensitive to the chosen bandwidth.

An additional caveat when using the RDD method is that the specification may be sensitive to the functional form used in modeling the relationship between the eligibility score and the outcome of interest. In the examples presented in this chapter, we assumed that the relation between the eligibility index and the outcome was linear. In reality, the relation could be more complex, including nonlinear relationships and interactions between variables. If you do not account for these complex relationships in the estimation, they might be mistaken for a discontinuity, leading to an incorrect interpretation of the RDD estimated impact. In practice, you can estimate program impact using various functional forms (linear, quadratic, cubic, quartic, and the like) to assess whether, in fact, the impact estimates are sensitive to functional form.

Finally, as discussed above, there are a few important conditions for the eligibility rule and cutoff. First, they must be unique to the program of interest. A poverty index ranking households or individuals, for example, may be used to target a variety of social programs to the poor. In this case, it will not be possible to isolate the impact of one particular antipoverty program from all the other programs that use the same targeting criteria. Second, the eligibility rule and cutoff should be resistant to manipulation by enumerators, potential beneficiaries, program administrators, or politicians. Manipulation of the eligibility index creates a discontinuity in the index that undermines the basic condition for the method to work: namely, that the eligibility index should be continuous around the cutoff.

Even with these limitations, RDD is a powerful impact evaluation method to generate unbiased estimates of a program's impact in the vicinity of the eligibility cutoff. The RDD takes advantage of the program assignment rules, using continuous eligibility indexes, which are already common in many social programs. When index-based targeting rules are applied, it is not necessary to exclude a group of eligible households or individuals from receiving the treatment for the sake of the evaluation because regression discontinuity design can be used instead.

Checklist: Regression Discontinuity Design

Regression discontinuity design requires that the eligibility index be continuous around the cutoff score and that units be similar in the vicinity above and below the cutoff score.

- ✓ Is the index continuous around the cutoff score at the time of the baseline?
- ✓ Is there any evidence of noncompliance with the rule that determines eligibility for treatment? Test whether all eligible units and no ineligible units have received the treatment. If you find noncompliance, you will need to combine RDD with an instrumental variable approach to correct for this "fuzzy discontinuity."[4]
- ✓ Is there any evidence that index scores may have been manipulated in order to influence who qualified for the program? Test whether the distribution of the index score is smooth at the cutoff point. If you find evidence of "bunching" of index scores either above or below the cutoff point, this might indicate manipulation.
- ✓ Is the cutoff unique to the program being evaluated, or is the cutoff used by other programs as well?

Additional Resources

- For accompanying material to the book and hyperlinks to additional resources, please see the Impact Evaluation in Practice website (http://www.worldbank.org/ieinpractice).
- For information about evaluating a cash transfer program using RDD, see the blog post on the World Bank Development Impact Blog (http://blogs.worldbank.org/impactevaluations/).
- For a review of practical issues in implementing RDD, see Imbens, Guido, and Thomas Lemieux. 2008. "Regression Discontinuity Designs: A Guide to Practice." *Journal of Econometrics* 142 (2): 615–35.

Notes

1. This is sometimes called a *proxy-means test*.
2. The continuous eligibility index is sometimes referred to as the forcing variable.
3. Technical note: Density was estimated using the univariate Epanechnikov kernel method.
4. In this case, you would use the location left or right of the cutoff point as an instrumental variable for actual program take-up in the first stage of a two-stage least-squares estimation.

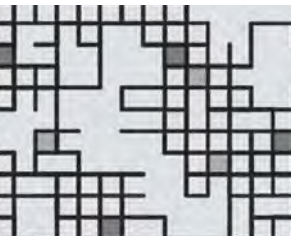
References

Barrera-Osorio, Felipe, Leigh Linden, and Miguel Urquiola. 2007. "The Effects of User Fee Reductions on Enrollment: Evidence from a Quasi-Experiment." Columbia University and World Bank, Washington, DC.

Duflo, Esther, Pascaline Dupas, and Michael Kremer. 2011. "Peer Effects, Teacher Incentives, and the Impact of Tracking: Evidence from a Randomized Evaluation in Kenya." *American Economic Review* 101: 1739–74.

Imbens, Guido, and Thomas Lemieux. 2008. "Regression Discontinuity Designs: A Guide to Practice." *Journal of Econometrics* 142 (2): 615–35.

Levy, Dan, and Jim Ohls. 2010. "Evaluation of Jamaica's PATH Conditional Cash Transfer Programme." *Journal of Development Effectiveness* 2 (4): 421–41.



CHAPTER 7

Difference-in-Differences

Evaluating a Program When the Rule of Assignment Is Less Clear

The three impact evaluation methods discussed up to this point—randomized assignment, instrumental variables (IV), and regression discontinuity design (RDD)—all produce estimates of the counterfactual through explicit program assignment rules that the evaluation team knows and understands. We have discussed why these methods offer credible estimates of the counterfactual with relatively few assumptions and conditions. The next two types of methods—difference-in-differences (DD) and matching methods—offer the evaluation team an additional set of tools that can be applied when the program assignment rules are less clear or when none of the three methods previously described is feasible. Both difference-in-differences and matching are commonly used in this case; however, both also typically require stronger assumptions than randomized assignment, IV, or RDD methods. Intuitively, if we do not know the program assignment rule, we have an additional unknown in our evaluation, about which we need to make assumptions. Since the assumptions we make are not necessarily true, using difference-in-differences or matching may not always provide reliable estimates of program impacts.

The Difference-in-Differences Method

Key Concept

Difference-in-differences compares the *changes* in outcomes over time between units that are enrolled in a program (the treatment group) and units that are not (the comparison group). This allows us to correct for any differences between the treatment and comparison groups that are constant over time.

The *difference-in-differences method* compares the *changes* in outcomes over time between a population that is enrolled in a program (the treatment group) and a population that is not (the comparison group). Take, for example, a road repair program that is carried out at the district level but cannot be randomly assigned between districts and is also not assigned based on an index with a clearly defined cutoff that would permit a regression discontinuity design. District boards can decide to enroll or not enroll in the program. One of the program's objectives is to improve access of the population to labor markets, and one of the outcome indicators is the employment rate. As discussed in chapter 3, simply observing the before-and-after change in employment rates for districts that enroll in the program will not capture the program's causal impact because many other factors are also likely to influence employment over time. At the same time, comparing districts that enrolled and did not enroll in the road repair program will be problematic if unobserved reasons exist for why some districts enrolled in the program and others did not (the selection bias problem discussed in the enrolled versus not-enrolled scenario).

However, what if we combined the two methods and compared the before-and-after changes in outcomes for a group that enrolled in the program with the before-and-after changes for a group that did not enroll in the program? The difference in the before-and-after outcomes for the enrolled group—*the first difference*—controls for factors that are constant over time in that group, since we are comparing the same group to itself. But we are still left with the factors that vary over time (*time-varying factors*) for this group. One way to capture those time-varying factors is to measure the before-and-after change in outcomes for a group that did not enroll in the program but was exposed to the same set of environmental conditions—*the second difference*. If we "clean" the first difference of other time-varying factors that affect the outcome of interest by subtracting the second difference, then we have eliminated a source of bias that worried us in the simple before-and-after comparisons. The difference-in-differences approach does what its name suggests. It combines the two counterfeit estimates of the counterfactual (before-and-after comparisons, and comparisons between those who choose to enroll and those who choose not to enroll) to produce a better estimate of the counterfactual. In the example of the road repair program, the DD method might compare the changes in employment before and after the program is implemented for individuals living in districts that enrolled in the program with the changes in employment in districts that did not enroll in the program.

It is important to note that what we are estimating here is the counterfactual for the *change* in outcomes for the treatment group: our estimate of this counterfactual is the change in outcomes for the comparison group. The treatment and comparison groups do not necessarily need to have the same conditions before the intervention. But for DD to be valid, the comparison group must accurately represent the change in outcomes that would have been experienced by the treatment group in the absence of treatment. To apply difference-in-differences, it is necessary to measure outcomes in the group that receives the program (the treatment group) and the group that does not (the comparison group), both before and after the program. In box 7.1, we present an example where the DD method was used to understand the impact of electoral incentives on implementation of a cash transfer program in Brazil and on school dropout rates.

Figure 7.1 illustrates the difference-in-differences method for the road repair example. Year 0 is the baseline year. In year 1, a treatment group of districts enrolls in the program, while a comparison group

Box 7.1: Using Difference-in-Differences to Understand the Impact of Electoral Incentives on School Dropout Rates in Brazil

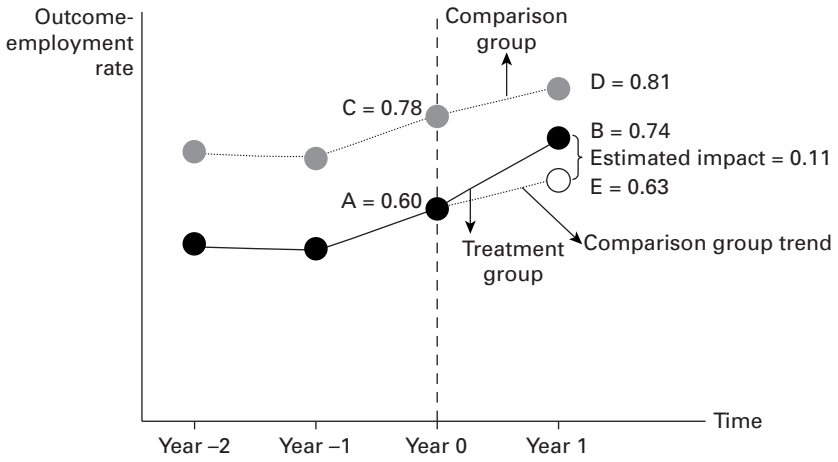
In an empirical study on local electoral incentives, De Janvry, Finan, and Sadoulet (2011) examined the impacts of a conditional cash transfer (CCT) in Brazil. The Bolsa Escola program gave mothers in poor households a monthly stipend conditional on their children's school attendance. The CCT was a federal program similar to Mexico's Oportunidades (see boxes 1.1 and 4.2), but it was decentralized to the municipal level. Municipal governments were responsible for identifying beneficiaries and implementing the program.

Using the difference-in-differences method, De Janvry, Finan, and Sadoulet estimated the impact of the program on school dropout rates. They found notable variation in the program's performance across municipalities. To explore this variation, the researchers compared the improvement in school dropout rates in municipalities with first-term versus second-term mayors. Their hypothesis was that, since Brazil has a two-term limit for local politicians, first-term mayors are concerned about reelection and therefore act differently than second-term mayors who do not have such concerns.

Overall, the program successfully reduced school dropout rates by an average of 8 percent for beneficiaries. The researchers found that the program's impact was 36 percent larger in municipalities with first-term mayors. Their conclusion was that reelection concerns incentivized local politicians to increase their effort in implementing the Bolsa Escola program.

Source: De Janvry, Finan, and Sadoulet 2011.

Figure 7.1 The Difference-in-Differences Method



Note: All differences between points should be read as vertical differences in outcomes on the vertical axis.

of districts is not enrolled. The outcome level (employment rate) for the treatment group goes from *A,* before the program starts, to *B* after the program has started, while the outcome for the comparison group goes from *C,* before the program started, to *D,* after the program has started.

You will remember our two counterfeit estimates of the counterfactual: the difference in outcomes before and after the intervention for the treatment group ($B - A$) and the difference in outcomes after the intervention between the treatment and comparison groups ($B - D$). In difference-in-differences, the estimate of the counterfactual is obtained by computing the change in outcomes for the comparison group ($D - C$), and then subtracting this from the change in outcomes for the treatment group ($B - A$). Using the change in outcomes for the comparison group as the estimate of the counterfactual for the change in outcomes for the treatment group is akin to assuming that, had the enrolled group not participated in the program, their outcome would have evolved over time along the same trend as the nonenrolled group: that is, the change in outcome for the enrolled group would have been from *A* to *E,* as shown in figure 7.1.

In summary, the impact of the program is simply computed as the difference between two differences:

$$\text{DD impact} = (B - A) - (D - C) = (0.74 - 0.60) - (0.81 - 0.78) = 0.11.$$

The relationships presented in figure 7.1 can also be presented in a simple table. Table 7.1 disentangles the components of the difference-in-differences estimates. The first row contains outcomes for the treatment group before the intervention (A) and after the intervention (B). The before-and-after comparison for the treatment group is the first difference ($B - A$). The second row contains outcomes for the comparison group before the intervention (C) and after the intervention (D), so the second difference is ($D - C$).

The difference-in-differences method computes the impact estimate as follows:

1. We calculate the difference in the outcome (Y) between the before and after situations for the treatment group ($B - A$).
2. We calculate the difference in the outcome (Y) between the before and after situations for the comparison group ($D - C$).
3. Then we calculate the difference between the difference in outcomes for the treatment group ($B - A$) and the difference for the comparison group ($D - C$), or difference-in-differences (DD) = ($B - A$) − ($D - C$). This difference-in-differences is our impact estimate.

We could also compute the difference-in-differences the other way across: first calculating the difference in the outcome between the treatment and the comparison group in the after situation, then calculating the difference in the outcome between the treatment and the comparison group in the before situation, and finally subtracting the latter from the former.

$$\text{DD impact} = (B - D) - (A - C) = (0.74 - 0.81) - (0.60 - 0.78) = 0.11.$$

Table 7.1 Calculating the Difference-in-Differences (DD) Method

	After	Before	Difference
Treatment/enrolled	B	A	$B - A$
Comparison/nonenrolled	D	C	$D - C$
Difference	$B - D$	$A - C$	DD = ($B - A$) − ($D - C$)

	After	Before	Difference
Treatment/enrolled	0.74	0.60	*0.14*
Comparison/nonenrolled	0.81	0.78	*0.03*
Difference	*−0.07*	*−0.18*	DD = *0.14 − 0.03 = 0.11*

How Is the Difference-in-Differences Method Helpful?

To understand how difference-in-differences is helpful, let us start with our second counterfeit estimate of the counterfactual discussed in chapter 3, which compared units that were enrolled in a program with those that were not enrolled in the program. Remember that the primary concern with this comparison was that the two sets of units may have had different characteristics and that it may be those characteristics—rather than the program—that explain the difference in outcomes between the two groups. The *unobserved* differences in characteristics were particularly worrying: by definition, it is impossible for us to include unobserved characteristics in the analysis.

The difference-in-differences method helps resolve this problem to the extent that many characteristics of units or individuals can reasonably be assumed to be constant over time (or *time-invariant*). Think, for example, of *observed* characteristics, such as a person's year of birth, a region's location close to the ocean, a town's climate, or a father's level of education. Most of these types of variables, although plausibly related to outcomes, will probably not change over the course of an evaluation. Using the same reasoning, we might conclude that many *unobserved* characteristics of individuals are also more or less constant over time. Consider, for example, personality traits or family health history. It might be plausible that these intrinsic characteristics of a person would not change over time.

Key Concept

Instead of comparing outcomes between the treatment and comparison groups after the intervention, the difference-in-differences methods compares *trends* between the treatment and comparison groups.

Instead of comparing outcomes between the treatment and comparison groups after the intervention, the difference-in-differences method compares *trends* between the treatment and comparison groups. The trend for an individual is the difference in outcome for that individual before and after the program. By subtracting the *before* outcome situation from the *after* situation, we cancel out the effect of all of the characteristics that are unique to that individual and that do not change over time. Interestingly, we are canceling out (or controlling for) not only the effect of *observed* time-invariant characteristics, but also the effect of *unobserved* time-invariant characteristics, such as those mentioned. Box 7.2 describes a study that used the difference-in-differences method to estimate the impact of increased police presence on incidences of car theft in Buenos Aires.

Box 7.2: Using Difference-in-Differences to Study the Effects of Police Deployment on Crime in Argentina

DiTella and Schargrodsky (2005) examined whether an increase in police forces reduced crime in Argentina. In 1994, a terrorist attack on a large Jewish center in Buenos Aires prompted the Argentine government to increase police protection for Jewish-affiliated buildings in the country.

Seeking to understand the impact of police presence on the incidence of crime, DiTella and Schargrodsky collected data on the number of car thefts per block in three neighborhoods in Buenos Aires before and after the terrorist attack. They then combined this information with geographic data on the location of Jewish-affiliated institutions in the neighborhoods. This study presented a different approach from typical crime regressions. Studies on the impact of policing often face an endogeneity problem, as governments tend to increase police presence in areas with higher crime rates. By contrast, the increase in police force deployment in Argentina was not related at all to the incidence of car thefts, so the study does not suffer this issue of simultaneous causality. DiTella and Schargrodsky were able to use the difference-in-differences method to estimate the impact of increased police presence on the incidence of car theft.

The results revealed a positive deterrent effect of police presence on crime; however, this effect was localized. In the blocks with Jewish-affiliated buildings that received police protection, car thefts decreased significantly compared with other blocks: by 75 percent. The researchers found no impacts on car thefts one or two blocks away from protected buildings.

Source: DiTella and Schargrodsky 2005.

The "Equal Trends" Assumption in Difference-in-Differences

Although difference-in-differences allows us to take care of differences between the treatment and comparison groups that are constant over time, it will not help us eliminate the differences between the treatment and comparison groups that change over time. In the example of the road repair program, if treatment areas also benefit from the construction of a new seaport at the same time as the road repair, we will not be able to separate out the effect from the road repair and from the seaport construction by using a difference-in-differences approach. For the method to provide a valid estimate of the counterfactual, we must assume that no such time-varying differences exist between the treatment and comparison groups.

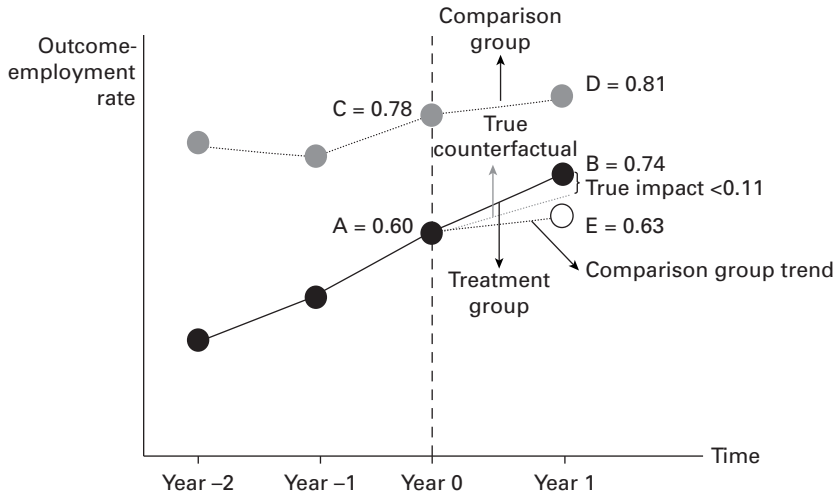
Another way to think about this is that in the absence of the program, the differences in outcomes between the treatment and comparison

groups would need to move in tandem. That is, without treatment, outcomes would need to increase or decrease at the same rate in both groups; we require that outcomes display *equal trends in the absence of treatment*.

Of course there is no way for us to prove that the differences between the treatment and comparison groups would have moved in tandem in the absence of the program. The reason is that we cannot observe what would have happened to the treatment group in the absence of the treatment—in other words, we cannot observe the counterfactual.

Thus when we use the difference-in-differences method, we must *assume* that, in the absence of the program, the outcome in the treatment group would have moved in tandem with the outcome in the comparison group. Figure 7.2 illustrates a violation of this fundamental assumption. If outcome trends are different for the treatment and comparison groups, then the estimated treatment effect obtained by difference-in-differences methods would be invalid, or biased. That's because the trend for the comparison group is not a valid estimate of the counterfactual trend that would have prevailed for the treatment group in the absence of the program. As shown in figure 7.2, if in reality outcomes for the comparison group grow more slowly than outcomes for the treatment group in the absence of the program, using the trend for the comparison group as an estimate of the counterfactual of the trend for the treatment group leads to a biased estimate of the program's impact; more specifically, we would overestimate the impact of the program.

Figure 7.2 Difference-in-Differences When Outcome Trends Differ



Testing the Validity of the "Equal Trends" Assumption in Difference-in-Differences

Even though it cannot be proved, the validity of the underlying assumption of equal trends can be assessed. A first validity check is to compare changes in outcomes for the treatment and comparison groups repeatedly before the program is implemented. In the road repair program, this means that we would compare the change in employment rate between treatment and comparison groups before the program starts: that is, between year −2 and year −1, and between year −1 and year 0. If the outcomes moved in tandem before the program started, we gain confidence that outcomes would have continued to move in tandem after the intervention. To check for equality of pre-intervention trends, we need at least two serial observations on the treatment and comparison groups before the start of the program. This means that the evaluation would require three serial observations: two pre-intervention observations to assess the preprogram trends, and at least one postintervention observation to assess impact with the difference-in-differences method.

A second way to test the assumption of equal trends would be to perform what is known as a *placebo test*. For this test, you perform an additional difference-in-differences estimation using a "fake" treatment group: that is, a group that you know was not affected by the program. Say, for example, that you estimate how additional tutoring for seventh-grade students affects their probability of attending school, and you choose eighth-grade students as the comparison group. To test whether seventh and eighth graders have the same trends in terms of school attendance, you could test whether eighth graders and sixth graders have the same trends. You know that sixth graders are not affected by the program, so if you perform a difference-in-differences estimation using eighth-grade students as the comparison group and sixth-grade students as the fake treatment group, you *have to* find a zero impact. If you do not, then the impact that you find must come from some underlying difference in trends between sixth graders and eighth graders. This, in turn, casts doubt on whether seventh graders and eighth graders can be assumed to have equal trends in the absence of the program.

A third way to test the assumption of equal trends would be to perform the placebo test not only with a fake treatment group, but also with a fake outcome. In the tutoring example, you may want to test the validity of using the eighth-grade students as a comparison group by estimating the impact of the tutoring on an outcome that you know is not affected by it, such as the number of siblings that the students have. If your difference-in-differences estimation finds an impact of the tutoring on the number of siblings that the students have, then you know that your comparison group must be flawed.

A fourth way to test the assumption of equal trends would be to perform the difference-in-differences estimation using different comparison groups. In the tutoring example, you would first do the estimation using eighth-grade students as the comparison group, and then do a second estimation using sixth-grade students as the comparison group. If both groups are valid comparison groups, you would find that the estimated impact is approximately the same in both calculations. In boxes 7.3 and 7.4, we present two examples of a difference-in-differences evaluation that used a combination of these methods to test the assumption of equal trends.

Box 7.3: Testing the Assumption of Equal Trends: Water Privatization and Infant Mortality in Argentina

Galiani, Gertler, and Schargrodsky (2005) used the difference-in-differences method to address an important policy question: Does privatizing the provision of water services improve health outcomes and help alleviate poverty? During the 1990s, Argentina initiated one of the largest privatization campaigns ever, transferring local water companies to regulated private companies. The privatization process took place over a decade, with the largest number of privatizations occurring after 1995, and eventually reached about 30 percent of the country's municipalities and 60 percent of the population.

The evaluation took advantage of that variation in ownership status over time to determine the impact of privatization on under-age-five mortality. Before 1995, the rates of child mortality were declining at about the same pace throughout Argentina. After 1995, mortality rates declined faster in municipalities that had privatized their water services.

The researchers argued that, in this context, the equal trends assumption behind difference-in-differences is likely to hold true. In particular, they showed that no differences in child mortality trends are observed between the comparison and treatment municipalities before the privatization movement began. They also showed that the decision to privatize was uncorrelated with economic shocks or historical levels of child mortality. They checked the strength of their findings by carrying out a placebo test with a fake outcome: they distinguished those causes of child mortality that are related water conditions, such as infectious and parasitic diseases, from those that are unrelated to water conditions, such as accidents and congenital diseases. They then tested the impact of privatization of water services separately for the two subsets of mortality causes. They found that privatization of water services was correlated with reductions in deaths from infectious and parasitic diseases, but not correlated with reductions in deaths from causes such as accidents and congenital diseases.

In the end, the evaluation determined that child mortality fell about 8 percent in areas that privatized, and that the effect was largest, about 26 percent, in the poorest areas, where the expansion of the water network was the greatest. This study shed light on a number of important policy debates surrounding the

(continued)

Box 7.3: Testing the Assumption of Equal Trends: Water Privatization and Infant Mortality in Argentina *(continued)*

privatization of public services. The researchers concluded that in Argentina, the regulated private sector proved more successful than the public sector in improving indicators of access, service, and most significantly, child mortality.

Source: Galiani, Gertler, and Schargrodsky 2005.

Box 7.4: Testing the Assumption of Equal Trends: School Construction in Indonesia

Duflo (2001) analyzed the medium- and long-term impacts of a program to build schools in Indonesia on education and labor market outcomes. In 1973, Indonesia embarked on a large-scale primary school construction program and built more than 61,000 primary schools. To target students who had not previously enrolled in school, the government allocated the number of schools to be constructed in each district in proportion to the number of unenrolled students in the district. Duflo sought to evaluate the program's impact on educational attainment and wages. Exposure to the treatment was measured by the number of schools in the region, and the treatment and comparison cohorts were identified by the age when the program was launched. The treatment group was composed of men born after 1962, as they would have been young enough to benefit from the new primary schools that were constructed in 1974. The comparison group was composed of men born before 1962 who would have been too old to benefit from the program.

Duflo used the difference-in-differences method to estimate the impact of the program on average educational attainment and wages, comparing the differences in outcomes among high- and low-exposure districts. To show that this was a valid estimation method, she first needed to test the assumption of equal trends across districts. To test this, Duflo used a placebo test with a fake treatment group. She compared the cohort ages 18–24 in 1974 with the cohort ages 12–17. Since both of these cohorts were too old to benefit from the new program, changes in their educational attainment should not be systematically different across districts. The estimate from this difference-in-differences regression was near zero. This result implied that educational attainment did not increase more rapidly before the program started in areas that would eventually become high-exposure districts than in low-exposure districts. The placebo test also showed that the identification strategy of relying on age at the time of school construction would work.

The evaluation found positive results on the educational attainment and wages of students who had high exposure to the program, meaning those who were under the age of eight when the schools were built. For these students, each new school constructed per 1,000 children was associated with a gain of 0.12 to 0.19 years in educational attainment and an increase of 3.0 percent to 5.4 percent in wages. The program also increased the probability that a child would complete primary school by 12 percent.

Source: Duflo 2001.

Evaluating the Impact of HISP: Using Difference-in-Differences

Difference-in-differences can be used to evaluate our Health Insurance Subsidy Program (HISP). In this scenario, you have two rounds of data on two groups of households: one group that enrolled in the program, and another that did not. Remembering the case of the enrolled and nonenrolled groups, you realize that you cannot simply compare the average health expenditures of the two groups because of selection bias. Because you have data for two periods for each household in the sample, you can use those data to solve some of these challenges by comparing the change in health expenditures for the two groups, assuming that the change in the health expenditures of the nonenrolled group reflects what would have happened to the expenditures of the enrolled group in the absence of the program (see table 7.2). Note that it does not matter which way you calculate the double difference.

Next, you estimate the effect using regression analysis (table 7.3). Using a simple linear regression to compute the simple difference-in-differences estimate, you find that the program reduced household health expenditures by US$8.16. You then refine your analysis by adding additional control variables. In other words, you use a multivariate linear regression that takes into account a host of other factors, and you find the same reduction in household health expenditures.

Table 7.2 Evaluating HISP: Difference-in-Differences Comparison of Means

	After (follow-up)	Before (baseline)	Difference
Enrolled	7.84	14.49	–6.65
Nonenrolled	22.30	20.79	1.51
Difference			DD = –6.65 – 1.51 = –8.16

Note: The table presents mean household health expenditures (in dollars) for enrolled and nonenrolled households, before and after the introduction of HISP.

Table 7.3 Evaluating HISP: Difference-in-Differences with Regression Analysis

	Linear regression	Multivariate linear regression
Estimated impact on household health expenditures	–8.16** (0.32)	–8.16** (0.32)

Note: Standard errors are in parentheses. Significance level: ** = 1 percent.

HISP Question 6

A. What are the basic assumptions required to accept this result from difference-in-differences?

B. Based on the result from difference-in-differences, should HISP be scaled up nationally?

Limitations of the Difference-in-Differences Method

Even when trends are equal before the start of the intervention, bias in the difference-in-differences estimation may still appear and go undetected. That's because DD attributes to the intervention any differences in trends between the treatment and comparison groups that occur from the time intervention begins. If any other factors are present that affect the difference in trends between the two groups and they are not accounted for in multivariate regression, the estimation will be invalid or biased.

Let us say that you are trying to estimate the impact on rice production of subsidizing fertilizer and are doing this by measuring the rice production of subsidized (treatment) farmers and unsubsidized (comparison) farmers before and after the distribution of the subsidies. If in year 1 there is a drought that affects only subsidized farmers, then the difference-in-differences estimate will produce an invalid estimate of the impact of subsidizing fertilizer. In general, any factor that disproportionately affects one of the two groups, and does so at the same time that the treatment group receives the treatment—and is not taken into account in the regression—has the potential to invalidate or bias the estimate of the impact of the program. Difference-in-differences assumes that no such factor is present.

Checklist: Difference-in-Differences

Difference-in-differences assumes that outcome trends are similar in the comparison and treatment groups before the intervention and that the only factors explaining differences in outcomes between the two groups are constant over time, apart from the program itself.

- ✓ Would outcomes have moved in tandem in the treatment and comparison groups in the absence of the program? This can be assessed by using several falsification tests, such as the following: (1) Are the outcomes in

the treatment and comparison groups moving in tandem before the intervention? If two rounds of data are available before the start of the program, test to see if any difference in trends appears between the two groups. (2) How about fake outcomes that should not be affected by the program? Are they moving in tandem before and after the start of the intervention in the treatment and comparison groups?

- ✓ Perform the difference-in-differences analysis using several plausible comparison groups. You should obtain similar estimates of the impact of the program.
- ✓ Perform the difference-in-differences analysis using your chosen treatment and comparison groups and a fake outcome that should not be affected by the program. You should find zero impact of the program on that outcome.
- ✓ Perform the difference-in-differences analysis using your chosen outcome variable with two groups that you know were not affected by the program. You should find zero impact of the program.

Additional Resources

- For accompanying material to the book and hyperlinks to additional resources, please see the Impact Evaluation in Practice website (http://www.worldbank.org/ieinpractice).
- For more on the unspoken assumptions behind difference-in-differences, see the World Bank Development Impact Blog (http://blogs.worldbank.org/impactevaluations).

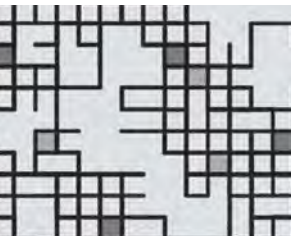
References

De Janvry, Alain, Frederico Finan, and Elisabeth Sadoulet. 2011. "Local Electoral Incentives and Decentralized Program Performance." *Review of Economics and Statistics* 94 (3): 672–85.

DiTella, Rafael, and Ernesto Schargrodsky. 2005. "Do Police Reduce Crime? Estimates Using the Allocation of Police Forces after a Terrorist Attack." *American Economic Review* 94 (1): 115–33.

Duflo, Esther. 2001. "Schooling and Labor Market Consequences of School Construction in Indonesia: Evidence from an Unusual Policy Experiment." *American Economic Review* 91 (4): 795–813.

Galiani, Sebastian, Paul Gertler, and Ernesto Schargrodsky. 2005. "Water for Life: The Impact of the Privatization of Water Services on Child Mortality." *Journal of Political Economy* 113 (1): 83–120.



CHAPTER 8

Matching

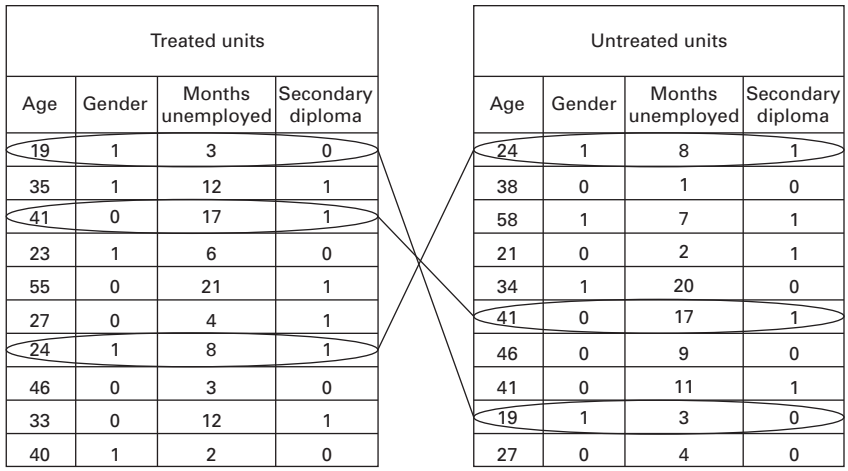
Constructing an Artificial Comparison Group

Key Concept
Matching uses large data sets and statistical techniques to construct the best possible comparison group based on observed characteristics.

The method described in this chapter consists of a set of statistical techniques that we will refer to collectively as *matching*. Matching methods can be applied in the context of almost any program assignment rules, as long as a group exists that has not participated in the program. Matching essentially uses statistical techniques to construct an artificial comparison group. For every possible unit under treatment, it attempts to find a nontreatment unit (or set of nontreatment units) that has the most similar characteristics possible. Consider a case in which you are attempting to evaluate the impact of a job training program on income and have a data set, such as income and tax records, that contains both individuals that enrolled in the program and individuals that did not enroll. The program that you are trying to evaluate does not have any clear assignment rules (such as randomized assignment or an eligibility index) that explain why some individuals enrolled in the program and others did not. In such a context, matching methods will enable you to identify the set of nonenrolled individuals that look most similar to the treated individuals, based on the characteristics that you have available in your data set. These matched nonenrolled individuals then become the comparison group that you use to estimate the counterfactual.

Finding a good match for each program participant requires approximating as closely as possible the characteristics that explain that individual's

Figure 8.1 Exact Matching on Four Characteristics



Treated units			
Age	Gender	Months unemployed	Secondary diploma
19	1	3	0
35	1	12	1
41	0	17	1
23	1	6	0
55	0	21	1
27	0	4	1
24	1	8	1
46	0	3	0
33	0	12	1
40	1	2	0

Untreated units			
Age	Gender	Months unemployed	Secondary diploma
24	1	8	1
38	0	1	0
58	1	7	1
21	0	2	1
34	1	20	0
41	0	17	1
46	0	9	0
41	0	11	1
19	1	3	0
27	0	4	0

decision to enroll in the program. Unfortunately, this is easier said than done. If the list of relevant observed characteristics is very large, or if each characteristic takes on many values, it may be hard to identify a match for each of the units in the treatment group. As you increase the number of characteristics or dimensions against which you want to match units that enrolled in the program, you may run into what is called the *curse of dimensionality*. For example, if you use only three important characteristics to identify the matched comparison group, such as age, gender, and whether the individual has a secondary school diploma, you will probably find matches for all participants enrolled in the program in the pool of those who are not enrolled (the nonenrolled), but you run the risk of leaving out other potentially important characteristics. However, if you increase the list of characteristics—say, to include number of children, number of years of education, number of months unemployed, number of years of experience, and so forth—your database may not contain a good match for most of the program participants who are enrolled, unless it contains a very large number of observations. Figure 8.1 illustrates matching based on four characteristics: age, gender, months unemployed, and secondary school diploma.

Propensity Score Matching

Fortunately, the curse of dimensionality can be quite easily solved using a method called *propensity score matching* (Rosenbaum and Rubin 1983). In this approach, we no longer need to try to match each enrolled unit to a

nonenrolled unit that has exactly the same value for all observed control characteristics. Instead, for each unit in the treatment group and in the pool of nonenrolled, we compute the *probability* that this unit will enroll in the program (the so-called propensity score) based on the observed values of its characteristics (the explanatory variables). This score is a real number between 0 and 1 that summarizes the influence of all of the observed characteristics on the likelihood of enrolling in the program. We should use only *baseline* observed characteristics to calculate the propensity score. This is because posttreatment characteristics might have been affected by the program itself, and using such characteristics to identify the matched comparison group would bias the results. When the treatment affects individual characteristics and we use those to match, we choose a comparison group that looks similar to the treated group because of the treatment itself. Without the treatment, those characteristics would look more different. This violates the basic requirement for a good estimate of the counterfactual: the comparison group must be similar in all aspects, except for the fact that the treatment group receives the treatment and the comparison group does not.

Once the propensity score has been computed for all units, then units in the treatment group can be matched with units in the pool of nonenrolled that have the closest propensity score.[1] These closest units become the comparison group and are used to produce an estimate of the counterfactual. The propensity score–matching method tries to mimic the randomized assignment to treatment and comparison groups by choosing for the comparison group those units that have similar propensities to the units in the treatment group. Since propensity score matching is not a randomized assignment method but tries to imitate one, it belongs to the category of quasi-experimental methods.

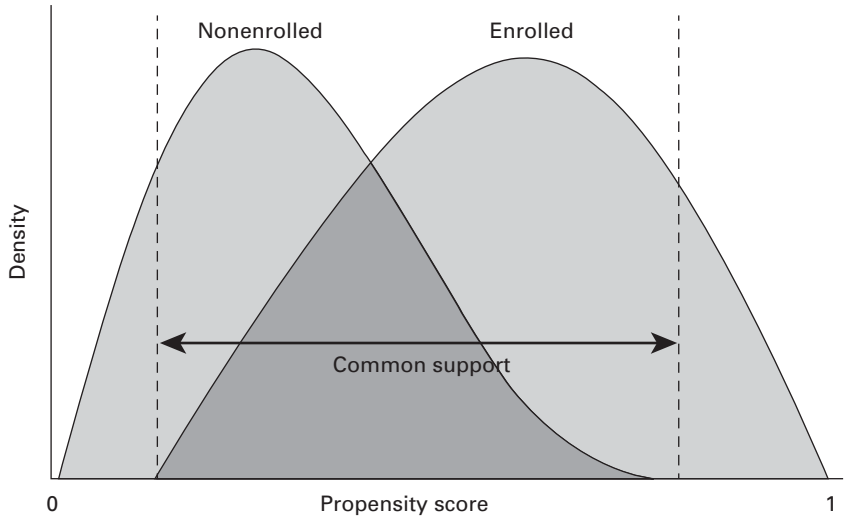
The average difference in outcomes between the treatment or enrolled units and their matched comparison units produces the estimated impact of the program. In summary, the program's impact is estimated by comparing the average outcomes of a treatment or enrolled group and the average outcomes among a statistically matched subgroup of units, the match being based on observed characteristics available in the data at hand.

For propensity score matching to produce estimates of a program's impact for all treated observations, each treatment or enrolled unit needs to be successfully matched to a nonenrolled unit.[2] In practice, however, it may be the case that for some enrolled units, no units in the pool of nonenrolled have similar propensity scores. In technical terms, there may be a *lack of common support*, or lack of overlap, between the propensity scores of the treatment or enrolled group and those of the pool of nonenrolled.

Figure 8.2 provides an example of lack of common support. First, we estimate the likelihood that each unit in the sample enrolls in the program based on the observed characteristics of that unit: that is, the propensity score. The figure shows the distribution of propensity scores separately for enrolled and nonenrolled. The issue is that these distributions do not overlap perfectly. In the middle of the distribution, matches are relatively easy to find because there are both the enrolled and nonenrolled with these levels of propensity scores. However, enrollees with propensity scores close to 1 cannot be matched to any nonenrolled because there are no nonenrolled with such high propensity scores. Intuitively, units that are highly likely to enroll in the program are so dissimilar to nonenrolling units that we cannot find a good match for them. Similarly, nonenrolled with propensity scores close to 0 cannot be matched to any enrollees because there are no enrollees with such low propensity scores. A lack of common support thus appears at the extremes, or tails, of the distribution of propensity scores. In this case, the matching procedure estimates the local average treatment effect (LATE) for observations on the common support.

The steps to be taken when applying propensity score matching are summarized in Jalan and Ravallion (2003).[3] First, you will need representative and highly comparable surveys in which it is possible to identify the units that enrolled in the program and those that did not. Second, you pool the two samples and estimate the probability that each individual enrolls in

Figure 8.2 Propensity Score Matching and Common Support



the program, based on individual characteristics observed in the survey. This step yields the propensity score. Third, you restrict the sample to units for which common support appears in the propensity score distribution. Fourth, for each enrolled unit, you locate a subgroup of nonenrolled units that have similar propensity scores. Fifth, you compare the outcomes for the treatment or enrolled units and their matched comparison or nonenrolled units. The difference in average outcomes for these two subgroups is the measure of the impact that can be attributed to the program for that particular treated observation. Sixth, the mean of these individual impacts yields an estimate of the local average treatment effect. In practice, commonly used statistical programs include preprogrammed commands that run steps 2 through 6 automatically.

Overall, it is important to remember three crucial issues about matching. First, matching methods can use only *observed* characteristics to construct a comparison group, since unobserved characteristics cannot be taken into account. If there are any unobserved characteristics that affect whether a unit enrolls in the program and also affect the outcome, then the impact estimates obtained with the matched comparison group would be biased. For a matching result to be unbiased, it requires the strong assumption that there are no unobserved differences in the treatment and comparison groups that are also associated with the outcomes of interest.

Second, matching must be done using only characteristics that are not affected by the program. Most characteristics that are measured after the start of the program would not fall into that category. If baseline (preintervention) data are not available and the only data are from after the intervention has started, the only characteristics we will be able to use to construct a matched sample are those (usually few) characteristics that are unaffected by a program, such as age and gender. Even though we would like to match on a much richer set of characteristics, including the outcomes of interest, we cannot do so because those are potentially affected by the intervention. Matching solely based on postintervention characteristics is not recommended. If baseline data are available, we can match based on a richer set of characteristics, including the outcomes of interest. Given that the data are collected before the intervention, those preintervention variables cannot have been affected by the program. However, if baseline data on outcomes are available, you should not use the matching method by itself. You should combine it with difference-in-differences to reduce the risk of bias. This procedure is discussed in the next section.

Third, the matching method's estimation results are only as good as the characteristics that are used for matching. While it is important to be

able to match using a large number of characteristics, even more important is to be able to match on the basis of characteristics that determine enrollment. The more we understand about the criteria used for participant selection, the better we will be able to construct the matched comparison group.

Combining Matching with Other Methods

Although the matching technique requires a significant amount of data and carries a significant risk of bias, it has been used to evaluate development programs in a wide array of settings. The most convincing uses of matching are those that combine matching with other methods and those that use the synthetic control method. In this section, we will discuss matched difference-in-differences and the synthetic control method.

Matched Difference-in-Differences

When baseline data on outcomes are available, matching can be combined with difference-in-differences to reduce the risk of bias in the estimation. As discussed, simple propensity score matching cannot account for unobserved characteristics that might explain why a group chooses to enroll in a program and that might also affect outcomes. Matching combined with difference-in-differences at least takes care of any unobserved characteristics that are constant across time between the two groups. It is implemented as follows:

1. Perform matching based on observed baseline characteristics (as discussed).
2. For each enrolled unit, compute the change in outcomes between the before and after periods (first difference).
3. For each enrolled unit, compute the change in outcomes between the before and after periods for this unit's matched comparison (second difference).
4. Subtract the second difference from the first difference; that is, apply the difference-in-differences method.
5. Finally, average out those double differences.

Boxes 8.1 and 8.2 provide examples of evaluations that used the matched difference-in-differences method in practice.

Box 8.1: Matched Difference-in-Differences: Rural Roads and Local Market Development in Vietnam

Mu and Van de Walle (2011) used propensity score matching in combination with difference-in-differences to estimate the impact of a rural road program on local market development at the commune level. From 1997 to 2001, the Vietnamese government rehabilitated 5,000 kilometers of rural roads. The roads were selected according to cost and population density criteria.

Since the communes that benefited from the rehabilitated roads were not randomly selected, the researchers used propensity score matching to construct a comparison group. Using data from a baseline survey, the researchers found a variety of factors at the commune level that influenced whether a road in the commune was selected for the program, such as population size, share of ethnic minorities, living standards, density of existing roads, and presence of passenger transport. They estimated propensity scores based on these characteristics and limited the sample size to the area of common support. This yielded 94 treatment and 95 comparison communes. To further limit the potential selection bias, the researchers used difference-in-differences to estimate the change in local market conditions.

Two years after the program, the results indicated that the road rehabilitation led to significant positive impacts on the presence and frequency of local markets and the availability of services. New markets developed in 10 percent more treatment communes than comparison communes. In treatment communes, it was more common for households to switch from agricultural to more service-related activities such as tailoring and hairdressing. However, the results varied substantially across communes. In poorer communes, the impacts tended to be higher due to lower levels of initial market development. The researchers concluded that small road improvement projects can have larger impacts if targeted at areas with an initially low market development.

Source: Mu and Van de Walle 2011.

Box 8.2: Matched Difference-in-Differences: Cement Floors, Child Health, and Maternal Happiness in Mexico

The Piso Firme program in Mexico offers households with dirt floors up to 50 square meters of concrete flooring (see box 2.1). Piso Firme began as a local program in the state of Coahuila, but was adopted nationally. Cattaneo and others (2009) took advantage of the geographic variation to evaluate the impact of this large-scale housing improvement effort on health and welfare outcomes.

The researchers used the difference-in-differences method in conjunction with matching to compare households in Coahuila with similar families in the neighboring state

(continued)

Box 8.2: Matched Difference-in-Differences: Cement Floors, Child Health, and Maternal Happiness in Mexico *(continued)*

of Durango, which at the time of the survey had not yet implemented the program. To improve comparability between the treatment and comparison groups, the researchers limited their sample to households in the neighboring cities that lie just on either side of the border between the two states. Within this sample, they used matching techniques to select treatment and comparison blocks that were the most similar. The pretreatment characteristics they used were the proportion of households with dirt floors, number of young children, and number of households within each block.

In addition to matching, the researchers used instrumental variables to recover the local average treatment effect from the intent-to-treat effect. With the offer of a cement floor as an instrumental variable for actually having cement floors, they found that the program led to an 18.2 percent reduction in the presence of parasites, a 12.4 percent reduction in the prevalence of diarrhea, and a 19.4 percent reduction in the prevalence of anemia. Furthermore, they were able to use variability in the amount of total floor space actually covered by cement to predict that a complete replacement of dirt floors with cement floors in a household would lead to a 78 percent reduction in parasitic infestations, a 49 percent reduction in diarrhea, an 81 percent reduction in anemia, and a 36 percent to 96 percent improvement in child cognitive development. The authors also collected data on adult welfare and found that cement floors make mothers happier, with a 59 percent increase in self-reported satisfaction with housing, a 69 percent increase in self-reported satisfaction with quality of life, a 52 percent reduction on a depression assessment scale, and a 45 percent reduction on a perceived stress assessment scale.

Cattaneo and others (2009) concluded by illustrating that Piso Firme has a larger absolute impact on child cognitive development at a lower cost than Mexico's large-scale conditional cash transfer program, Oportunidades/Progresa, as well as comparable programs in nutritional supplementation and early childhood cognitive stimulation. The cement floors also prevented more parasitic infections than the common deworming treatment. The authors state that programs to replace dirt floors with cement floors are likely to improve child health cost-effectively in similar contexts.

Source: Cattaneo and others 2009.

The Synthetic Control Method

The synthetic control method allows for impact estimation in settings where a single unit (such as a country, a firm, or a hospital) receives an intervention or is exposed to an event. Instead of comparing this treated unit to a group of untreated units, the method uses information about the characteristics of the treated unit and the untreated units to construct a "synthetic," or artificial, comparison unit by weighting each untreated unit in such a way that the synthetic comparison unit most closely resembles the

treated unit. This requires a long series of observations over time of the characteristics of both the treated unit and the untreated units. This combination of comparison units into a synthetic unit provides a better comparison for the treated unit than any untreated unit individually. Box 8.3 provides an example of an evaluation that used the synthetic control method.

Box 8.3: The Synthetic Control Method: The Economic Effects of a Terrorist Conflict in Spain

Abadie and Gardeazabal (2003) used the synthetic control method to investigate the economic effects of the terrorist conflict in the Basque Country. In the early 1970s, the Basque Country was one of the richest regions in Spain; however, by the late 1990s, after 30 years of conflict, it had dropped to the sixth position in per capita gross domestic product (GDP). At the onset of terrorism in the early 1970s, the Basque Country differed from other Spanish regions in characteristics that are thought to be related to potential for economic growth. Therefore a comparison of GDP growth between the Basque economy and the rest of Spain would reflect both the effect of terrorism and the effect of these differences in economic growth determinants before the onset of terrorism. In other words, the difference-in-differences approach would yield biased results of the impact of terrorism on economic growth in the Basque Country. To deal with this situation, the authors used a combination of other Spanish regions to construct a "synthetic" comparison region.

Source: Abadie and Gardeazabal 2003.

Evaluating the Impact of HISP: Using Matching Techniques

Having learned about matching techniques, you may wonder whether you could use them to estimate the impact of the Health Insurance Subsidy Program (HISP). You decide to use some matching techniques to select a group of nonenrolled households that look similar to the enrolled households based on baseline observed characteristics. To do this, you use your statistical software's matching package. First, it estimates the probability that a household will enroll in the program based on the observed values of characteristics (the explanatory variables), such as the age of the household head and of the spouse, their level of education, whether the head of the household is a female, whether the household is indigenous, and so on.

We will carry out matching using two scenarios. In the first scenario, there is a large set of variables to predict enrollment, including socioeconomic household characteristics. In the second scenario, there is little information to predict enrollment (only education and age of the

household head). As shown in table 8.1, the likelihood that a household is enrolled in the program is smaller if the household is older, more educated, headed by a female, has a bathroom, or owns larger amounts of land. By contrast, being indigenous, having more household members, having a dirt floor, and being located further from a hospital all increase the likelihood that a household is enrolled in the program. So overall, it seems that poorer and less-educated households are more likely to be enrolled, which is good news for a program that targets poor people.

Now that the software has estimated the probability that each household is enrolled in the program (the propensity score), you check the distribution of the propensity score for the enrolled and matched comparison households. Figure 8.3 shows that common support (when using the full set of explanatory variables) extends across the whole distribution of the propensity score. In fact, none of the enrolled households fall outside the area of common support. In other words, we are able to find a matched comparison household for each of the enrolled households.

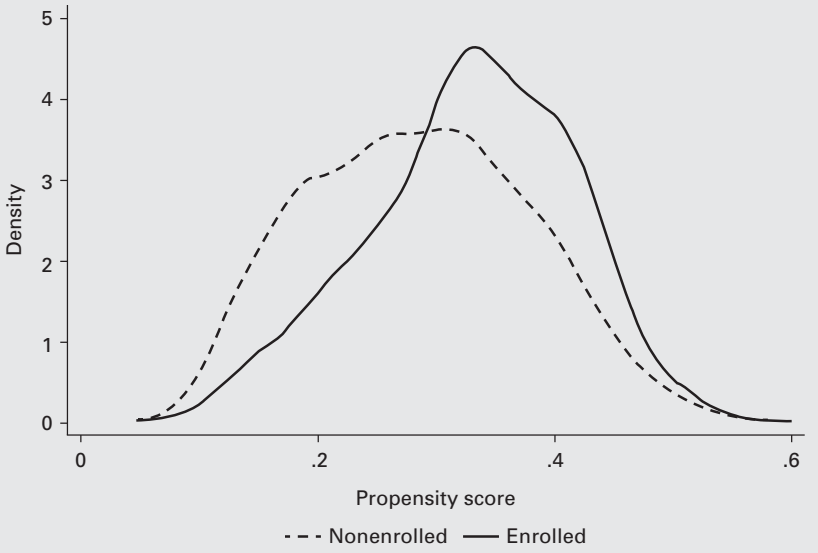
Table 8.1 Estimating the Propensity Score Based on Baseline Observed Characteristics

Dependent variable: Enrolled = 1	Full set of explanatory variables	Limited set of explanatory variables
Explanatory variables: Baseline observed characteristics	**Coefficient**	**Coefficient**
Head of household's age (years)	−0.013**	−0.021**
Spouse's age (years)	−0.008**	−0.041**
Head of household's education (years)	−0.022**	
Spouse's education (years)	−0.016*	
Head of household is female =1	−0.020	
Indigenous = 1	0.161**	
Number of household members	0.119**	
Dirt floor = 1	0.376**	
Bathroom = 1	−0.124**	
Hectares of land	−0.028**	
Distance to hospital (km)	0.002**	
Constant	−0.497**	0.554**

Note: Probit regression. The dependent variable is 1 if the household enrolled in HISP, and 0 otherwise. The coefficients represent the contribution of each listed explanatory variable to the probability that a household enrolled in HISP.

Significance level: * = 5 percent, ** = 1 percent.

Figure 8.3 Matching for HISP: Common Support



You decide to use *nearest neighbor matching*; that is, you tell the software to locate, for each enrolled household, the nonenrolled household that has the closest propensity score to the enrolled household. The software now restricts the sample to those households in the enrolled and nonenrolled groups for which it can find a match in the other group.

To obtain the estimated impact using the matching method, you first compute the impact for each enrolled household individually (using each household's matched comparison household), and then average those individual impacts. Table 8.2 shows that the impact estimated from applying this procedure is a reduction of US$9.95 in household health expenditures.

Finally, the software also allows you to compute the standard error on the estimated impact using linear regression (table 8.3).[4]

You realize that you also have information on baseline outcomes in your survey data, so you decide to carry out matched difference-in-differences in addition to using the full set of explanatory variables. That is, you compute the difference in household health expenditures at follow-up between enrolled and matched comparison households; you compute the difference in household health expenditures at baseline between enrolled and matched comparison households; and then you compute the difference between these two differences. Table 8.4 shows the result of this matched difference-in-differences approach.

Table 8.2 Evaluating HISP: Matching on Baseline Characteristics and Comparison of Means

	Enrolled	Matched comparison	Difference
Household health expenditures (US$)	7.84	17.79 (using full set of explanatory variables)	–9.95
		19.9 (using limited set of explanatory variables)	–11.35

Note: This table compares mean household health expenditures for enrolled households and matched comparison households.

Table 8.3 Evaluating HISP: Matching on Baseline Characteristics and Regression Analysis

	Linear regression (Matching on full set of explanatory variables)	Linear regression (Matching on limited set of explanatory variables)
Estimated impact on household health expenditures (US$)	–9.95** (0.24)	–11.35** (0.22)

Note: Standard errors are in parentheses. Significance level: ** = 1 percent.

Table 8.4 Evaluating HISP: Difference-in-Differences Combined with Matching on Baseline Characteristics

		Enrolled	Matched comparison using full set of explanatory variables	Difference
Household health expenditures (US$)	Follow-up	7.84	17.79	–9.95
	Baseline	14.49	15.03	0.54
				Matched difference-in-differences = –9.41**
				(0.19)

Note: Standard error is in parentheses and was calculated using linear regression. Significance level: ** = 1 percent.

HISP Question 7

A. What are the basic assumptions required to accept these results based on the matching method?
B. Why are the results from the matching method different if you use the full versus the limited set of explanatory variables?
C. What happens when you compare the result from the matching method with the result from randomized assignment? Why do you think the results are so different for matching on a limited set of explanatory variables? Why is the result more similar when matching on a full set of explanatory variables?
D. Based on the result from the matching method, should HISP be scaled up nationally?

Limitations of the Matching Method

Although matching procedures can be applied in many settings, regardless of a program's assignment rules, they have several serious shortcomings. First, they require extensive data sets on large samples of units, and even when those are available, there may be a lack of common support between the treatment or enrolled group and the pool of nonparticipants. Second, matching can only be performed based on observed characteristics; by definition, we cannot incorporate unobserved characteristics in the calculation of the propensity score. So for the matching procedure to identify a valid comparison group, we must be sure that no systematic differences in unobserved characteristics between the treatment units and the matched comparison units exist[5] that could influence the outcome (Y). Since we cannot *prove* that there are no such unobserved characteristics that affect both participation and outcomes, we must *assume* that none exist. This is usually a very strong assumption. Although matching helps control for *observed* background characteristics, we can never rule out bias that stems from *unobserved* characteristics. In summary, the assumption that no selection bias has occurred stemming from unobserved characteristics is very strong, and most problematically, it cannot be tested.

Matching alone is generally less robust than the other evaluation methods we have discussed, since it requires the strong assumption that there are no unobserved characteristics that simultaneously affect program participation and outcomes. Randomized assignment, instrumental variable, and regression discontinuity design, on the other hand, do not require the untestable assumption that there are no such unobserved variables.

They also do not require such large samples or as extensive background characteristics as propensity score matching.

In practice, matching methods are typically used when randomized assignment, instrumental variable, and regression discontinuity design options are not possible. So-called *ex post matching* is very risky when no baseline data are available on the outcome of interest or on background characteristics. If an evaluation uses survey data that were collected after the start of the program (that is, ex post) to infer what people's background characteristics were at baseline, and then matches the treated group to a comparison group using those inferred characteristics, it may inadvertently match based on characteristics that were also affected by the program; in that case, the estimation result would be invalid or biased.

By contrast, when baseline data are available, matching based on baseline background characteristics can be very useful when it is combined with other techniques, such as difference-in-differences, which allows us to correct for differences between the groups that are fixed over time. Matching is also more reliable when the program assignment rule and underlying variables are known, in which case matching can be performed on those variables.

By now, it is probably clear that impact evaluations are best designed before a program begins to be implemented. Once the program has started, if one has no way to influence how it is allocated and no baseline data have been collected, few, if any, rigorous options for the impact evaluation will be available.

Checklist: Matching

Matching relies on the assumption that enrolled and nonenrolled units are similar in terms of any unobserved variables that could affect both the probability of participating in the program and the outcome.

- ✓ Is program participation determined by variables that cannot be observed? This cannot be directly tested, so you will need to rely on theory, common sense, and good knowledge of the setting of the impact evaluation for guidance.
- ✓ Are the observed characteristics well balanced between matched subgroups? Compare the observed characteristics of each treatment and its matched comparison group of units at baseline.
- ✓ Can a matched comparison unit be found for each treatment unit? Check whether sufficient common support exists in the distribution of the propensity scores. Small areas of common support indicate that enrolled and nonenrolled persons are very different, and that casts doubt as to whether matching is a credible method.

Additional Resources

- For accompanying material for this book and hyperlinks to additional resources, please see the Impact Evaluation in Practice website (http://www.worldbank.org/ieinpractice).
- For more information on matching, see Rosenbaum, Paul. 2002. *Observational Studies*, second edition. Springer Series in Statistics. New York: Springer-Verlag.
- For more on implementing propensity score matching, see Heinrich, Carolyn, Alessandro Maffioli, and Gonzalo Vásquez. 2010. "A Primer for Applying Propensity-Score Matching. Impact-Evaluation Guidelines." Technical Note IDB-TN-161, Inter-American Development Bank, Washington, DC.

Notes

1. Technical note: In practice, many definitions of what constitutes the closest or nearest unit are used to perform matching. The nearest comparison units can be defined based on a stratification of the propensity score—the identification of the treatment unit's nearest neighbors, based on distance, within a given radius—or using kernel techniques. It is considered good practice to check the robustness of matching results by using various matching algorithms. See Rosenbaum (2002) for more details.
2. The discussion on matching in this book focuses on one-to-one matching. Various other types of matching, such as one-to-many matching or replacement/nonreplacement matching, will not be discussed. In all cases, however, the conceptual framework described here would still apply.
3. A detailed review of matching can be found in Rosenbaum (2002).
4. Technical note: When the enrolled units' propensity scores are not fully covered by the area of common support, standard errors should be estimated using bootstrapping rather than linear regression.
5. For readers with a background in econometrics, this means that participation is independent of outcomes, given the background characteristics used to do the matching.

References

Abadie, Alberto, and Javier Gardeazabal. 2003. "The Economic Costs of Conflict: A Case Study of the Basque Country." *American Economic Review* 93 (1): 113–32.

Cattaneo, Matias D., Sebastian Galiani, Paul J. Gertler, Sebastian Martinez, and Rocio Titiunik. 2009. "Housing, Health, and Happiness." *American Economic Journal: Economic Policy* 1 (1): 75–105.

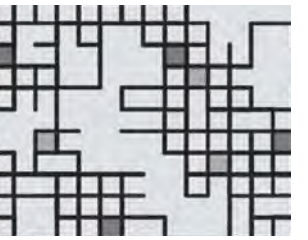
Heinrich, Carolyn, Alessandro Maffioli, and Gonzalo Vásquez. 2010. "A Primer for Applying Propensity-Score Matching. Impact-Evaluation Guidelines." Technical Note IDB-TN-161, Inter-American Development Bank, Washington, DC.

Jalan, Jyotsna, and Martin Ravallion. 2003. "Estimating the Benefit Incidence of an Antipoverty Program by Propensity-Score Matching." *Journal of Business & Economic Statistics* 21 (1): 19–30.

Mu, Ren, and Dominique Van de Walle. 2011. "Rural Roads and Local Market Development in Vietnam." *Journal of Development Studies* 47 (5): 709–34.

Rosenbaum, Paul. 2002. *Observational Studies*, second edition. Springer Series in Statistics. New York: Springer-Verlag.

Rosenbaum, Paul, and Donald Rubin. 1983. "The Central Role of the Propensity Score in Observational Studies of Causal Effects." *Biometrika* 70 (1): 41–55.



CHAPTER 9

Addressing Methodological Challenges

Heterogeneous Treatment Effects

We have seen that most impact evaluation methods produce valid estimates of the counterfactual only under specific assumptions. The main risk in applying any given method is that its underlying assumptions do not hold true, resulting in biased estimates of the program's impact. But there are also a number of other risks that are common to most of the methodologies that we have discussed. We will discuss the key ones in this chapter.

One type of risk arises if you are estimating a program's impact on an entire group, and your results mask some differences in responses to the treatment among different recipients, that is, heterogeneous treatment effects. Most impact evaluation methods assume that a program affects outcomes in a simple, linear way for all the units in the population.

If you think that different subpopulations may have experienced the impact of a program very differently, then you may want to consider having stratified samples by each subpopulation. Say, for example, that you are interested in knowing the impact of a school meal program on girls, but only 10 percent of the students are girls. In that case, even a large random sample of students may not contain a sufficient number of girls to allow you to estimate the impact of the program on girls. For your

evaluation's sample design, you would want to stratify the sample on the basis of gender and include a sufficiently large number of girls to allow you to detect a given effect size.

Unintended Behavioral Effects

When conducting an impact evaluation, you may also induce unintended behavioral responses from the population that you are studying, as follows:

- The *Hawthorne effect* occurs when the mere fact that you are observing units makes them behave differently (see box 9.1).
- The *John Henry effect* happens when comparison units work harder to compensate for not being offered a treatment (see box 9.1).
- *Anticipation* can lead to another type of unintended behavioral effect. In a randomized rollout, units in the comparison group may expect to receive the program in the future and begin changing their behavior before the program actually reaches them.

Box 9.1: Folk Tales of Impact Evaluation: The Hawthorne Effect and the John Henry Effect

The term *Hawthorne effect* refers to experiments that were carried out from 1924 to 1932 at the Hawthorne Works, an electric equipment factory in the U.S. state of Illinois. The experiments tested the impact of changing working conditions (such as increasing or decreasing the intensity of light) on workers' productivity, and they found that any changes in working conditions (more or less light, more or fewer breaks, and the like) led to an increase in productivity. This was interpreted as an observation effect: workers who were part of the experiment saw themselves as special, and their productivity increased because of this and not because of the change in working conditions. While the original experiments later became the subject of controversy and were somewhat discredited, the term Hawthorne effect stuck.

The term *John Henry effect* was coined by Gary Saretsky in 1972 to refer to legendary American folk hero John Henry, a "steel-driving man" tasked with driving a steel drill into rock to make holes for explosives during construction of a railroad tunnel. According to legend, when he learned that he was being compared to a steam drill, he worked much harder so as to outperform the machine. Alas, he died as a result. But the term lives on to denote how comparison units sometimes work harder to compensate for not being offered a treatment.

Sources: Landsberger 1958; Levitt and List 2009; Saretsky 1972.

- *Substitution bias* is another behavioral effect that affects the comparison group: units that were not selected to receive the program may be able to find good substitutes through their own initiative.

Behavioral responses that disproportionately affect the comparison group are an issue because they may undermine the internal validity of the evaluation results, even if you use randomized assignment as the evaluation method. A comparison group that works harder to compensate for not being offered a treatment, or that changes behavior anticipating the program, is not a good representation of the counterfactual.

If you have reason to believe that these unintended behavioral responses may be present, then building in additional comparison groups that are completely unaffected by the intervention is sometimes an option—one that allows you to explicitly test for such responses. It might also be a good idea to gather qualitative data in order to better understand behavioral responses.

Imperfect Compliance

Imperfect compliance is a discrepancy between assigned treatment status and actual treatment status. Imperfect compliance happens when some units assigned to the treatment group do not receive treatment, and when some units assigned to the comparison group receive treatment. In chapter 5, we discussed imperfect compliance in reference to randomized assignment, but imperfect compliance can also occur in regression discontinuity design (as discussed in chapter 6) and in difference-in-differences (chapter 7). Before you can interpret the impact estimates produced by any method, you need to know whether imperfect compliance has occurred in the program.

Imperfect compliance can occur in a variety of ways:

- Not all intended program participants actually participate in the program. Sometimes units that are assigned to a program choose not to participate.
- Some intended participants are excluded from the program because of administrative or implementation errors.
- Some units of the comparison group are mistakenly offered the program and enroll in it.
- Some units of the comparison group manage to participate in the program, even though it is not offered to them.
- The program is assigned based on a continuous eligibility index, but the eligibility cutoff is not strictly enforced.

- *Selective migration* takes place based on treatment status. For example, the evaluation may compare outcomes for treated and nontreated municipalities, but individuals may choose to move to another municipality if they do not like the treatment status of their municipality.

In general, in the presence of imperfect compliance, standard impact evaluation methods produce intention-to-treat estimates. However, the local average treatment effect can be recovered from the intention-to-treat estimates using the instrumental variable approach.

In chapter 5, we presented the intuition for dealing with imperfect compliance in the context of randomized assignment. Using an adjustment for the percentage of compliers in the evaluation sample, we were able to recover the local average treatment effect for the compliers from the intention-to-treat estimate. This "fix" can be extended to other methods through application of the more general instrumental variable approach. The instrumental variable contains an external source of variation that helps you clear up, or correct, the bias that may stem from imperfect compliance. In the case of randomized assignment with imperfect compliance, we used a 0/1 variable (a so-called *dummy* variable) that takes the value 1 if the unit was originally assigned to the treatment group, and 0 if the unit was originally assigned to the comparison group. During the analysis stage, the instrumental variable is used in the context of a *two-stage regression* that allows you to identify the impact of the treatment on the compliers.

The logic of the instrumental variable approach can be extended in the context of other evaluation methods:

- In the context of regression discontinuity design, the instrumental variable you would use is a 0/1 variable that indicates whether a unit is located on the ineligible side or the eligible side of the cutoff score.
- In the context of selective migration, a possible instrumental variable for the location of the individual after the start of the program would be the location of the individual before the announcement of the program.

Despite the possibility of addressing imperfect compliance using instrumental variables, three points are important to remember:

1. From a technical point of view, it is not desirable to have a large portion of the comparison group enroll in the program. As the portion of the comparison group that enrolls in the program increases, the fraction of compliers in the population will decrease, and the local average treatment effect estimated with the instrumental variable method will be valid only for a shrinking fraction of the population of interest. If this

continues too long, the results may lose all policy significance, since they would no longer be applicable to a sufficient portion of the population of interest.

2. Conversely, it is not desirable to have a large portion of the treatment group remain unenrolled. Again, as the portion of the treatment group that enrolls in the program decreases, the fraction of compliers in the population decreases. The local average treatment effect estimated with the instrumental variable method will be valid only for a shrinking fraction of the population of interest.
3. As discussed in chapter 5, the instrumental variable method is valid only under certain circumstances; it is definitely not a universal solution.

Spillovers

Spillovers (or spillover effects) are another common issue that may affect evaluations, whether they use the randomized assignment, regression discontinuity design, or difference-in-differences method. A *spillover* happens when an intervention affects a nonparticipant, and it might be positive or negative. There are four types of spillover effects, according to Angelucci and Di Maro (2015):

- *Externalities.* These are effects that go from treated subjects to untreated subjects. For example, vaccinating the children in a village against influenza decreases the probability that nonvaccinated inhabitants of the same village will catch this disease. This is an example of a positive externality. Externalities may also be negative. For example, a farmer's crop could be partially destroyed when his neighbor applies an herbicide on his own plot and some of the herbicide blows to the other side of the property line.
- *Social interactions.* Spillovers might result from social and economic interactions between treated and nontreated populations, leading to indirect impacts on the nontreated. For example, a student who receives a tablet as part of a learning enhancement program shares the tablet with another student who does not participate in the program.
- *Context equilibrium effects.* These effects happen when an intervention affects the behavioral or social norms within the given context, such as a treated locality. For example, increasing the amount of resources that treated health centers receive so they can extend their range of services might affect the expectations from the population about what should be the range of services offered at all health centers.

- *General equilibrium effects.* These effects happen when interventions affect the supply and demand for good or services, and thereby change the market price for those services. For example, a program that gives poor women vouchers to use private facilities for childbirth might suddenly increase the demand for services at private facilities, thereby increasing the price of the service for everyone else. Box 9.2 presents an example of negative spillovers due to general equilibrium effects in the context of a job training program.

If the nonparticipant who experiences the spillover is a member of the comparison group, then the spillover violates the basic requirement that the outcome of one unit should be unaffected by the particular assignment of treatments to other units. This *stable unit treatment value assumption* (SUTVA) is necessary to ensure that randomized assignment yields unbiased estimates of impact. Intuitively, if the comparison group is indirectly affected by the treatment received by the treatment group (for example, comparison students borrow tablets from treated students), then the comparison does

Box 9.2: Negative Spillovers Due to General Equilibrium Effects: Job Placement Assistance and Labor Market Outcomes in France

Job placement assistance programs are popular in many industrialized countries. Governments contract with a third-party entity to assist unemployed workers in their job search. Many studies find that these counseling programs have a significant and positive impact on job seekers.

Crépon and others (2013) investigated whether giving job assistance to young, educated job seekers in France might have negative effects on other job seekers who were not supported by the program. They hypothesized that a spillover mechanism might be at work: when the labor market is not growing much, helping one job seeker to find a job might come at the detriment of another job seeker who might otherwise have gotten the job that the counseled job seeker obtained. To investigate this hypothesis, they carried out a randomized experiment that included 235 labor markets (such as cities) in France. These labor markets were randomly allocated to one of five groups, which varied in terms of the proportion of job seekers to be assigned to counseling treatment (0 percent, 25 percent, 50 percent, 75 percent, and 100 percent). Within each labor market, eligible job seekers were randomly assigned to the treatment following this proportion. After eight months, the researchers found that unemployed youths who were assigned to the program were significantly more likely to have found a stable job than those who were not. But these gains appear to have come partly at the expense of eligible workers who did not benefit from the program.

Source: Crépon and others 2013.

not accurately represent what would have happened to the treatment group in absence of the treatment (the counterfactual).

If the nonparticipant who experiences the spillover is not a member of the comparison group, then the SUTVA assumption would hold, and the comparison group would still provide a good estimate of the counterfactual. However, we still would want to measure the spillover because it represents a real impact of the program. In other words, comparing the outcomes of the treatment and comparison groups would yield unbiased estimates of the impact of the treatment on the treated group, but this would not take into account the impact of the program on *other* groups.

A classic example of spillovers due to externalities is presented by Kremer and Miguel (2004), who examined the impact of administering deworming medicine to children in Kenyan schools (box 9.3). Intestinal worms are parasites that can be transmitted from one person to another through contact with contaminated fecal matter. When a child receives deworming medicine, her worm load will decrease, but so will the worm load of people living in the same environment, as they will no longer come in contact with the child's worms. Thus in the Kenya example, when the medicine was administered to the children in one school, it benefited not only those children (a direct benefit) but also those in neighboring schools (an indirect benefit).

As depicted in figure 9.1, deworming children in group A schools also diminishes the number of worms that affect children who don't attend group A schools. In particular, it may diminish the number of worms that affect children who attend group B comparison schools, which are located close to group A schools. However, comparison schools farther away from group A schools—the so-called group C schools—do not experience such spillover effects because the medicine administered in group A does not kill any of the worms that affect children attending group C schools. The evaluation and its results are discussed in more detail in box 9.3.

Designing an Impact Evaluation That Accounts for Spillovers

Say that you are designing an impact evaluation for a program where you think it's likely that spillovers will occur. How should you approach this? The first thing to do is to realize that the objective of your evaluation needs to be expanded. While a standard evaluation aims to estimate the impact (or causal effect) of a program on an outcome of interest for units receiving the treatment, an evaluation with spillovers will have to answer two questions:

1. *The standard evaluation question for the direct impact*. What is the impact (or causal effect) of a program on an outcome of interest for units receiving the treatment? This is the direct impact that the program has on treated groups.

Box 9.3: Working with Spillovers: Deworming, Externalities, and Education in Kenya

The Primary School Deworming Project in Busia, Kenya, was designed to test a variety of aspects of worm treatment and prevention. It was carried out by the Dutch nonprofit International Child Support Africa, in cooperation with the ministry of health. The project involved 75 schools with a total enrollment of more than 30,000 students between the ages of 6 and 18. The students were treated with worm medication in accordance with World Health Organization recommendations and also received worm prevention education in the form of health lectures, wall charts, and teacher training.

Due to administrative and financial constraints, the rollout was phased according to the alphabetical order of the school's name, with the first group of 25 schools starting in 1998, the second group in 1999, and the third group in 2001. By randomizing at the level of school, Kremer and Miguel (2004) were able both to estimate the impact of deworming on a school and to identify spillovers across schools using exogenous variation in the closeness of comparison schools to treatment schools. Although compliance with the randomized design was relatively high (with 75 percent of students assigned to the treatment receiving worm medication, and only a small percentage of the comparison group units receiving treatment), the researchers were also able to take advantage of noncompliance to determine within-school health externalities, or spillovers.

The direct effect of the interventions was to reduce moderate-to-heavy worm infections by 26 percentage points for students who took the worm medication. Meanwhile, moderate-to-heavy infections among students who attended treatment schools but did not take the medication fell by 12 percentage points through an indirect spillover effect. There were also externalities between schools.

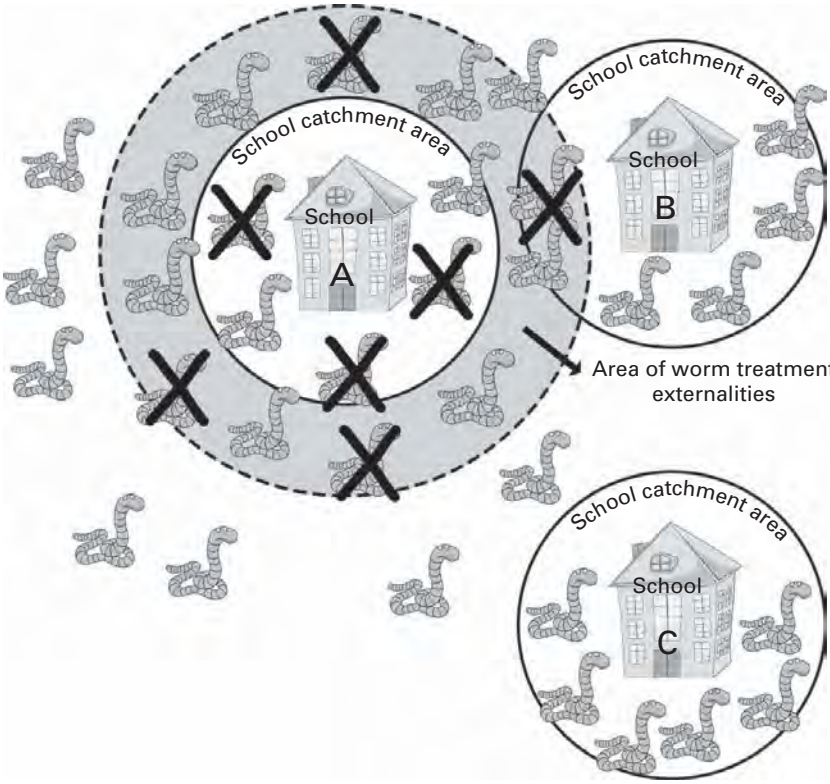
Because the cost of worm treatment is so low and the health and education effects relatively high, the researchers concluded that deworming is a relatively cost-efficient way to improve participation rates in schools. The study also illustrates that tropical diseases such as worms may play a significant role in educational outcomes and strengthens claims that Africa's high disease burden may be contributing to its low income. Thus Kremer and Miguel argue that the study makes a strong case for public subsidies for disease treatments with similar spillover benefits in developing countries.

Source: Kremer and Miguel 2004.

2. *An additional evaluation question for the indirect impact.* What is the impact (or causal effect) of a program on an outcome of interest for units *not* receiving the treatment? This is the indirect impact that the program has on nontreated groups.

To estimate the direct impact on treated groups, you will need to choose the comparison group in such a way that it is not affected by spillovers. For example, you might require that the treatment and comparison villages,

Figure 9.1 A Classic Example of Spillovers: Positive Externalities from Deworming School Children



clinics, or households be located sufficiently far from one another that spillovers are unlikely.

To estimate the indirect impact on nontreated groups, you should identify an additional comparison group for each nontreated group that may be affected by spillovers. For instance, community health workers may undertake household visits to provide information to parents about the benefits of improved dietary diversity for their children. Let us assume that the community health workers visit only some households in any given village. You may be interested in spillover effects on children in nonvisited households, in which case you would need to find a comparison group for these children. At the same time, it may be possible that the intervention also affects adults' dietary diversity. If such an indirect effect is of interest to the evaluation, a comparison group would also be needed among adults. As the number of potential spillover channels increases, the design can quickly become rather complicated.

Evaluations with spillovers pose some specific challenges. First, when spillover effects are likely, it is important to understand the mechanism of spillover: biological, social, environmental, or the like. If we don't know what the spillover mechanism is, we will be unable to accurately choose comparison groups that are and are not affected by spillovers. Second, an evaluation with spillovers requires more extensive data collection than one where this is not a concern: there is an additional comparison group (in the preceding example, nearby villages). You may also need to collect data on additional units (in the preceding example, adults in households targeted by nutrition visits for children). Box 9.4 examines how researchers handled spillovers in an evaluation of a conditional cash transfer program in Mexico.

Box 9.4: Evaluating Spillover Effects: Conditional Cash Transfers and Spillovers in Mexico

Angelucci and De Giorgi (2009) examined spillovers in Mexico's Progresa program, which provided conditional cash transfers to households (see boxes 1.1 and 4.2). The researchers sought to explore whether there was risk sharing within villages. If households shared risk, then eligible households could be transferring part of the cash transfer to ineligible households through loans or gifts.

The Progresa program was phased in over two years, with 320 villages randomly selected to receive the cash transfers in 1998, and 186 in 1999. So between 1998 and 1999 there were 320 treatment villages and 186 comparison villages. Within the treatment villages, a household's eligibility for Progresa transfers was determined based on poverty status, and census data were collected for both groups. This created four subgroups within the sample: eligible and ineligible populations within both treatment and comparison villages. Assuming that the program did not indirectly affect comparison villages, the ineligible households in the comparison villages provided a valid counterfactual for the ineligible households in the treatment villages, for the purpose of estimating within-village spillovers to ineligible households.

The researchers found evidence of positive spillovers on consumption. Adult food consumption increased about 10 percent per month in ineligible households in treatment villages. This was about half the average increase in food consumption among eligible households. The results also supported the hypothesis of risk-sharing in villages. Ineligible households in treatment villages received more loans and transfers from family and friends than did ineligible households in comparison villages. This implies that the spillover effect operated through local insurance and credit markets.

Based on these results, Angelucci and De Giorgi concluded that previous evaluations of Progresa underestimated the impact of the program by 12 percent because they did not account for indirect effects on ineligible households within treatment villages.

Source: Angelucci and De Giorgi 2009.

Attrition

Attrition bias is another common issue that may affect evaluations, whether they use the randomized assignment, regression discontinuity design, or difference-in-differences methods. *Attrition* occurs when parts of the sample disappear over time, and researchers are not able to find all initial members of the treatment and comparison groups in follow-up surveys or data. For example, of the 2,500 households surveyed in the baseline, researchers are able to find only 2,300 in a follow-up survey two years later. If researchers go back and attempt to resurvey the same group, say, 10 years later, they might be able to find even fewer original households.

Attrition might happen for various reasons. For example, members of households or even entire households might move to another village, city, region, or even country. In a recent example of a long-term follow-up of an early childhood development intervention in Jamaica, at the 22-year follow-up survey, 18 percent of the sample had migrated abroad (see box 9.5). In other cases, respondents might no longer be willing to respond to an additional survey. Or conflicts and lack of security in the area might prevent the research team from carrying out a survey in some locations that were included in the baseline.

Attrition can be problematic for two reasons. First, the follow-up sample might no longer accurately represent the population of interest. Remember that when we choose the sample at the time of the randomized assignment, we choose it so that it accurately represents the population of interest. In other words, we choose a sample that has external validity for our population of interest. If the follow-up survey or data collection is marred by substantial attrition, we would be concerned that the follow-up sample might represent only a specific subset of the population of interest. For example, if the most educated people in the original sample are also the ones who migrate, our follow-up survey would miss those educated people and no longer accurately represent the population of interest, which included those educated people.

Second, the follow-up sample might no longer be balanced between the treatment and comparison group. Say you are trying to evaluate a program that tries to boost girls' education, and that educated girls are more likely to move to the city to look for work. Then your follow-up survey might show disproportionately high attrition in the treatment group, compared with the comparison group. This could affect the internal validity of the program: by comparing the treatment and comparison units that you find at follow-up, you will no longer be able to accurately estimate the impact of the program.

Box 9.5: Attrition in Studies with Long-Term Follow-Up: Early Childhood Development and Migration in Jamaica

Attrition can be especially problematic where many years have passed between the baseline and follow-up surveys. In 1986, a team at the University of the West Indies began a study to measure long-term outcomes from an early childhood intervention in Jamaica. In 2008, a follow-up was conducted when the original participants were 22 years old. It was challenging to track down all of the original study participants.

The intervention was a two-year program that provided psychosocial stimulation and food supplementation to growth-stunted toddlers in Kingston, Jamaica. A total of 129 children were randomly assigned to one of three treatment arms or a comparison group. The researchers also surveyed 84 nonstunted children for a second comparison group. In the follow-up, the researchers were able to resurvey about 80 percent of the participants. There was no evidence of selective attrition in the whole sample, meaning that there were no significant differences in the baseline characteristics of those who could be surveyed at 22 years, compared with those who could not be surveyed. However, when considering the subgroup of children who had become migrant workers, there were signs of selective attrition. Out of 23 migrant workers, nine had dropped out of the sample, and a significantly larger share of these belonged to the treatment group. This implied that the treatment was associated with migration. Since migrant workers typically earned more than those who remained in Jamaica, this made it difficult to estimate impacts.

To address the potential bias from attrition among migrant workers, the researchers used econometric techniques. They predicted earnings for the migrant workers that had dropped out of the sample through an ordinary least squares (OLS) regression using treatment status, gender, and migration as determinants. Using these predictions in the impact estimation, the researchers found that the program had impressive results. The early childhood intervention increased earnings by 25 percent for the treatment group. This effect was large enough for the stunted treatment group to catch up to the nonstunted comparison group 20 years later.

Source: Gertler and others 2014; Grantham-McGregor and others 1991.

If you find attrition during a follow-up survey, the following two tests can help you assess the extent of the problem. First, check whether the baseline characteristics of the units that dropped out of the sample are statistically equal to baseline characteristics of the units that were successfully resurveyed. As long as the baseline characteristics of both groups are not

statistically different, your new sample should continue to represent the population of interest.

Second, check whether the attrition rate in the treatment group is similar to the attrition rate in the comparison group. If the attrition rates are significantly different, then there is a concern that your sample is no longer valid and you may need to use various statistical techniques to try to correct this. One common method is *inverse probability weighting*, a method that statistically reweights the data (in this case, the follow-up data) so as to correct for the fact that a portion of the original respondents is missing. The method reweighs the follow-up sample so it looks similar to the baseline sample.[1]

Timing and Persistence of Effects

The likely channels of transmission between inputs, activities, outputs, and outcomes might happen immediately, soon, or with a substantial time lag, and are usually closely related to changes in human behavior. Chapter 2 emphasized how important it is to think about these channels and plan before the intervention starts, and to develop a clear causal chain for the program being evaluated. For the sake of simplicity, we have been abstracting from timing issues. But it is important to consider aspects related to timing when designing an evaluation.

First, programs do not necessarily become fully effective immediately after they start (King and Behrman 2009). Program administrators may need time to get a program running, beneficiaries may not immediately benefit because behavioral changes take time, and institutions may not immediately adjust their behavior either. On the other hand, once institutions and beneficiaries change certain behaviors, it might be the case that they continue even if the program is discontinued. For example, a program that incentivizes households to sort and recycle garbage and save energy might continue to be effective after incentives are removed, if it manages to change household norms about how to handle garbage and energy. When designing an evaluation, you need to be very careful (and realistic) in assessing how long it might take for a program to reach full effectiveness. It might be necessary to carry out several follow-up surveys to gauge the impact of the program over time, or even after the program is discontinued. Box 9.6 illustrates an evaluation where some effects only became apparent after the initial intervention was discontinued.

Box 9.6: Evaluating Long-Term Effects: Subsidies and Adoption of Insecticide-Treated Bed Nets in Kenya

Dupas (2014) designed an impact evaluation to measure both the short- and long-term impacts on demand for insecticide-treated bed nets (ITNs) in Busia, Kenya. Using a two-phase pricing experiment, Dupas randomly assigned households to various subsidy levels for a new type of ITN. One year later, all households in a subset of villages were given the opportunity to purchase the same net. This allowed researchers to measure households' willingness to pay for the ITNs and how it changed depending on the subsidy given in the first phase of the program.

Overall, the results indicated that a one-time subsidy had significantly positive impacts on adoption of ITNs and willingness to pay in the longer term. In the first phase of the experiment, Dupas found that households that received a subsidy that decreased the price of the ITN from US$3.80 to US$0.75 were 60 percent more likely to purchase it. When the ITN was offered for free, the adoption rate increased to 98 percent. In the long run, the higher adoption rates translated to a higher willingness to pay, as households saw the benefits of having an ITN. Those that received one of the larger subsidies in the first phase were three times more likely to purchase another ITN in the second phase at more than double the price.

The results from this study imply that a learning effect occurs in ITN interventions. This suggests that it is important to consider the impacts of interventions in the long run, as well to uncover the persistence of effects.

Source: Dupas 2014.

Additional Resources

- For accompanying material to the book and hyperlinks to additional resources, please see the Impact Evaluation in Practice website (http://www.worldbank.org/ieinpractice).

Note

1. A more advanced statistical method would be to estimate "sharp bounds" on treatment effects (see Lee 2009).

References

Angelucci, Manuela, and Giacomo De Giorgi. 2009. "Indirect Effects of an Aid Program: How Do Cash Transfers Affect Ineligibles' Consumption." *American Economic Review* 99 (1): 486–508.

Angelucci, Manuela, and Vicenzo Di Maro. 2015. "Programme Evaluation and Spillover Effects." *Journal of Development Effectiveness*. doi: 10.1080/19439342.2015.1033441.

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

Dupas, Pascaline. 2014. "Short-Run Subsidies and Long-Run Adoption of New Health Products: Evidence from a Field Experiment." *Econometrica* 82 (1): 197–228.

Gertler, Paul, James Heckman, Rodrigo Pinto, Arianna Zanolini, Christel Vermeersch, Susan Walker, Susan M. Chang, and Sally Grantham-McGregor. 2014. "Labor Market Returns to an Early Childhood Stimulation Intervention in Jamaica." *Science* 344 (6187): 998–1001.

Grantham-McGregor, Sally, Christine Powell, Susan Walker, and John Himes. 1991. "Nutritional Supplementation, Psychosocial Stimulation and Development of Stunted Children: The Jamaican Study." *Lancet* 338: 1–5.

King, Elizabeth M., and Jere R. Behrman. 2009. "Timing and Duration of Exposure in Evaluations of Social Programs." *World Bank Research Observer* 24 (1): 55–82.

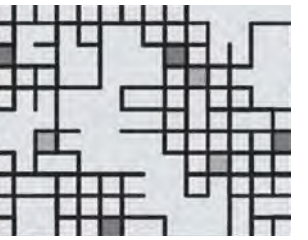
Kremer, Michael, and Edward Miguel. 2004. "Worms: Identifying Impacts on Education and Health in the Presence of Treatment Externalities." *Econometrica* 72 (1): 159–217.

Landsberger, Henry A. 1958. *Hawthorne Revisited*. Ithaca, NY: Cornell University Press.

Lee, David. 2009. "Training, Wages, and Sample Selection: Estimating Sharp Bounds on Treatment Effects." *Review of Economic Studies* 76 (3): 1071–102.

Levitt, Steven D., and John A. List. 2009. "Was There Really a Hawthorne Effect at the Hawthorne Plant? An Analysis of the Original Illumination Experiments." Working Paper 15016, National Bureau of Economic Research, Cambridge, MA.

Saretsky, Gary. 1972. "The OEO P.C. Experiment and the John Henry Effect." *Phi Delta Kappan* 53: 579–81.



CHAPTER 10

Evaluating Multifaceted Programs

Evaluating Programs That Combine Several Treatment Options

Up to now, we have discussed programs that include only one kind of treatment. In reality, many highly relevant policy questions arise in the context of multifaceted programs: that is, programs that combine several treatment options.[1] Policy makers may be interested in knowing not only whether or not a program works, but also whether the program works better than another or at lower cost. For example, if we want to increase school attendance, is it more effective to implement demand-side interventions (such as cash transfers to families) or supply-side interventions (such as greater incentives for teachers)? If we introduce the two interventions together, do they work better than each of them alone? In other words, are they complementary? Alternatively, if program cost-effectiveness is a priority, you may well want to determine the optimal level of services that the program should deliver. For instance, what is the optimal duration of a vocational training program? Does a six-month program have a greater effect on trainees' finding jobs than a three-month program? If so, is the difference large enough to justify the additional resources needed for a six-month program? Finally, policy makers may be interested in how to alter an existing program

to make it more effective, and they might want to test a variety of mechanisms in order to find which one(s) work best.

Beyond simply estimating the impact of an intervention on an outcome of interest, impact evaluations can help to answer broader questions such as these:

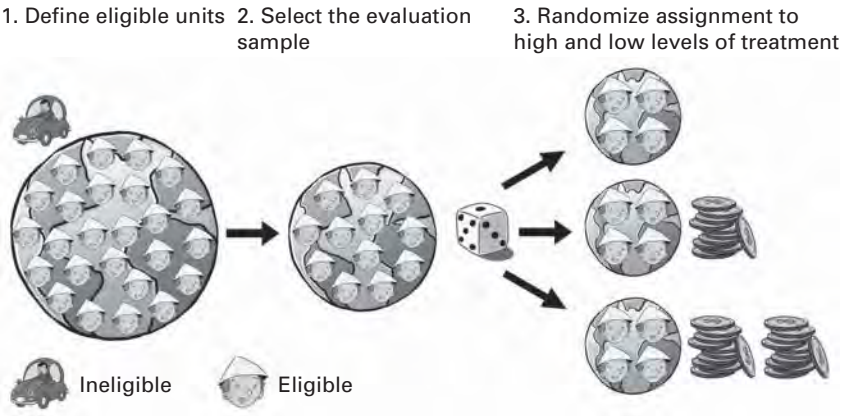
- What is the impact of one treatment compared with the impact of another treatment? For example, what is the impact on children's cognitive development of a program providing parenting training as opposed to a nutrition intervention?
- Is the joint impact of a first treatment and a second treatment larger than the sum of the two individual impacts? For example, is the total impact of the parenting intervention and the nutrition intervention greater than, less than, or equal to the sum of the effects of the two individual interventions?
- What is the additional impact of a higher-intensity treatment compared with a lower-intensity treatment? For example, what is the effect on the cognitive development of stunted children if a social worker visits them at home every two weeks, as compared with visiting them only once a month?

This chapter provides examples of how to design impact evaluations for two types of multifaceted programs: ones with multiple levels of the same treatment, and ones with multiple treatments. First, we discuss how to design an impact evaluation for a program with multiple treatment levels. Then we turn to how to disentangle the various kinds of impact of a program with multiple treatments. The discussion assumes that we are using the randomized assignment method, but it can be generalized to other methods.

Evaluating Programs with Varying Treatment Levels

It is relatively easy to design an impact evaluation for a program with varying treatment levels. Imagine that you are trying to evaluate the impact of a program that has two levels of treatment: high (for example, biweekly visits) and low (say, monthly visits). You want to evaluate the impact of both options, and you also want to know how much the additional visits affect outcomes. To do this, you can run a lottery to decide who receives the high level of treatment, who receives the low level of treatment, and who is assigned to the comparison group. Figure 10.1 illustrates this process.

Figure 10.1 Steps in Randomized Assignment of Two Levels of Treatment



As in standard randomized assignment, step 1 is to define the population of eligible units for your program. Step 2 is to select a random sample of units to be included in the evaluation, the *evaluation sample*. Once you have the evaluation sample, in step 3 you then randomly assign units to the group receiving high-level treatment, the group receiving low-level treatment, or the comparison group. As a result of randomized assignment to multiple treatment levels, you will have created three distinct groups:

- Group A constitutes the comparison group.
- Group B receives the low level of treatment.
- Group C receives the high level of treatment.

When correctly implemented, randomized assignment ensures that the three groups are similar. Therefore, you can estimate the impact of the high level of treatment by comparing the average outcome for group C with the average outcome for group A. You can also estimate the impact of the low level of treatment by comparing the average outcome for group B with that for group A. Finally, you can assess whether the high-level treatment has a larger impact than the low-level treatment by comparing the average outcomes for groups B and C.

Key Concept

When evaluating programs with *n* different treatment levels, there should be *n* treatment groups, plus a comparison group.

Estimating the impact of a program with more than two treatment levels will follow the same logic. If there are three levels of treatment, the randomized assignment process will create three different treatment groups, plus a comparison group. In general, with *n* different treatment levels, there will be *n* treatment groups, plus a comparison group. Box 10.1 and 10.2 provide examples of impact evaluations that test modalities of different intensity or multiple treatment options.

Box 10.1: Testing Program Intensity for Improving Adherence to Antiretroviral Treatment

Pop-Eleches and others (2011) used a multilevel cross-cutting design to evaluate the impact of using short message service (SMS) reminders on HIV/AIDS patients' adherence to antiretroviral therapy at a rural clinic in Kenya. The study varied the intensity of the treatment along two dimensions: how often the messages were sent to patients (daily or weekly), and the length of the messages (short or long). Short messages included only a reminder ("This is your reminder."), while long messages included a reminder as well as a word of encouragement ("This is your reminder. Be strong and courageous, we care about you."). A total of 531 patients were assigned to one of four treatment groups or the comparison group. The treatment groups were short weekly messages, long weekly messages, short daily messages, or long daily messages.

One-third of the sample was allocated to the control group, and the remaining two-thirds of the sample were allocated evenly to each of the four intervention groups. A sequence of random numbers between 0 and 1 was generated. Four equal intervals between 0 and 2/3 corresponded to the four intervention groups, while the value interval from 2/3 to 1 corresponded to the control group.

The investigators found that weekly messages increased the percentage of patients achieving 90 percent adherence to antiretroviral therapy by approximately 13–16 percent, compared with no messages. These weekly messages were also effective at reducing the frequency of treatment interruptions, which have been shown to be an important cause of treatment-resistant failure in resource-limited settings. Contrary to expectations, adding words of encouragement in the longer messages was not more effective than either a short message or no message.

The investigators also found that while weekly messages improved adherence, daily messages did not, but they were not able to distinguish as to why the weekly messages were most effective. It is possible that habituation, or the diminishing of a response to a frequently repeated stimulus, may explain this finding, or patients may simply have found the daily messages to be intrusive.

Table B10.1.1 Summary of Program Design

Group	Type of message	Frequency of message	Number of patients
1	Reminder only	Weekly	73
2	Reminder + encouragement	Weekly	74
3	Reminder only	Daily	70
4	Reminder + encouragement	Daily	72
5	None (comparison group)	None	139

Source: Pop-Eleches and others 2011.

Box 10.2: Testing Program Alternatives for Monitoring Corruption in Indonesia

In Indonesia, Olken (2007) used a cross-cutting design to test different methods for controlling corruption, from a top-down enforcement approach to more grassroots community monitoring. He used a randomized assignment methodology in more than 600 villages that were building roads as part of a nationwide infrastructure improvement project.

One of the multiple treatments included randomly selecting some villages to be informed that their construction project would be audited by a government agent. Then, to test community participation in monitoring, the researchers implemented two interventions. They passed out invitations to community accountability meetings, and they provided comment forms that could be submitted anonymously. To measure the levels of corruption, an independent team of engineers and surveyors took core samples of the new roads, estimated the cost of the materials used, and then compared their calculations to the reported budgets.

Olken found that increasing government audits (from about a 4 percent chance of being audited to a 100 percent chance) reduced missing expenditures by about 8 percentage points (from 24 percent). Increasing community participation in monitoring had an impact on missing labor but not on missing expenditures. The comment forms were effective only when they were distributed to children at school to give to their families and not when handed out by the village leaders.

Source: Olken 2007.

Evaluating Multiple Interventions

In addition to comparing various levels of treatment, you may want to compare entirely different treatment options. In fact, policy makers usually prefer comparing the relative merits of different interventions, rather than simply knowing the impact of only a single intervention.

Imagine that you want to evaluate the impact on school attendance of a program with two different interventions: cash transfers to the students' families that are conditional on school enrollment and free bus transportation to school. First, you may want to know the impact of each intervention separately. This case is virtually identical to the one where you test different levels of treatment of one intervention: instead of randomly assigning units to high and low levels of treatments and the comparison group, you could randomly assign them to a cash transfers group, a free bus transportation group, and the comparison group. In general, with n different interventions, there will be n treatment groups plus a comparison group.

Apart from wanting to know the impact of each intervention separately, you may also want to know whether the combination of the two is better

than just the sum of the individual effects. Seen from the participants' point of view, the program is available in three different forms: conditional cash transfers only, free bus transportation only, or a combination of conditional cash transfers and free bus transportation.

Randomized assignment for a program with two interventions is very much like the process for a program with a single intervention. The main difference is the need to conduct several independent lotteries instead of one. This produces a *crossover design*, sometimes called a cross-cutting design. Figure 10.2 illustrates this process. As before, step 1 is to define the population of units eligible for the program. Step 2 is to select a random sample of eligible units from the population to form the evaluation sample. Once you obtain the evaluation sample, step 3 is to randomly assign units from the evaluation sample to a treatment group and a comparison group. In step 4, you use a second lottery to randomly assign a subset of the treatment group to receive the second intervention. Finally, in step 5, you conduct another lottery to assign a subset of the initial comparison group to receive the second intervention, while the other subset will remain as a pure comparison.[2] As a result of the randomized assignment to the two treatments, you will have created four groups, as illustrated in figure 10.3.

- Group A receives both interventions (cash transfers and bus transportation).
- Group B receives intervention 1 but not intervention 2 (cash transfers only).
- Group C does not receive intervention 2 but receives intervention 1 (bus transportation only).
- Group D receives neither intervention 1 nor intervention 2 and constitutes the pure comparison group.

When correctly implemented, randomized assignment ensures that the four groups are similar. You can therefore estimate the impact of the first intervention by comparing the outcome (such as school attendance rate) for group B with the outcome for the pure comparison group, group D. You can also estimate the impact of the second intervention by comparing the outcome for group C to the outcome for the pure comparison group, group D. In addition, this design also makes it possible to compare the incremental impact of receiving the second intervention when a unit already receives the first one. Comparing the outcomes of group A and group B will yield the impact of the second intervention for those units that have already received the first intervention. Comparing the outcomes of group A and group C will yield the impact of the first intervention for those units that have already received the second intervention.

Figure 10.2 Steps in Randomized Assignment of Two Interventions

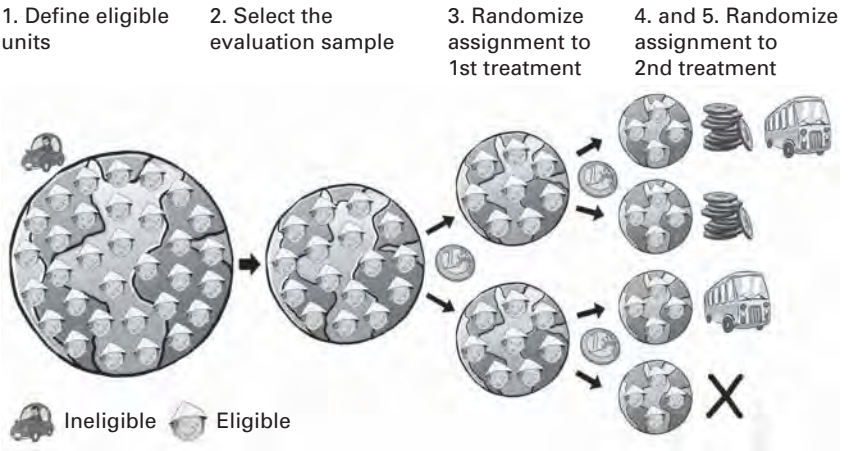
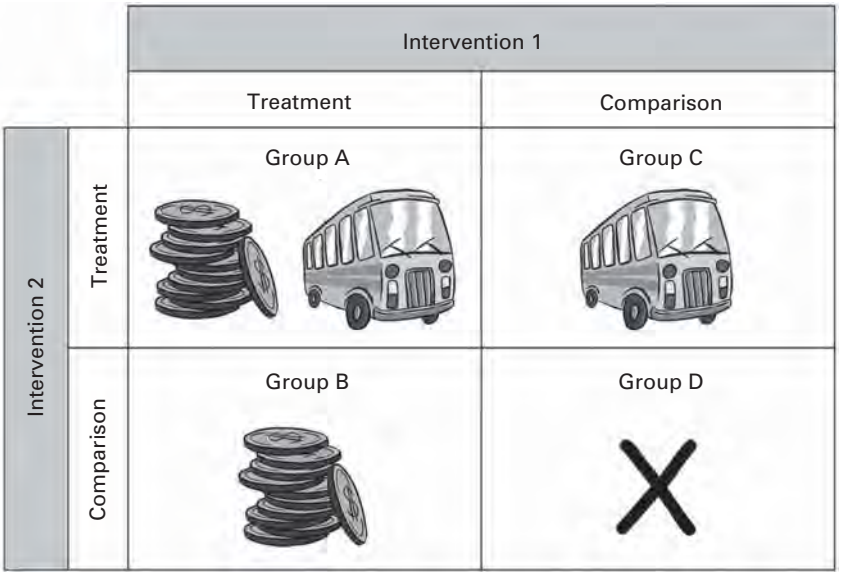


Figure 10.3 Crossover Design for a Program with Two Interventions



The foregoing description has used the example of randomized assignment to explain how an impact evaluation can be designed for a program with two different interventions. When a program comprises more than two interventions, the number of lotteries can be increased, and the evaluation can be further subdivided to construct groups that receive the various combinations of interventions. Designs with multiple treatments and multiple treatment levels can also be implemented. Even if the number of groups increases, the basic theory behind the design remains the same, as described earlier.

However, evaluating more than one or two interventions will create practical challenges both for the evaluation and for program operation, as the complexity of the design will increase exponentially with the number of treatment arms. To evaluate one intervention, only two groups are needed: one treatment group and one comparison group. To evaluate two interventions, four groups are needed: three treatment groups and one comparison group. If you were to evaluate three interventions, including all possible combinations among the three interventions, you would need $2 \times 2 \times 2 = 8$ groups in the evaluation. In general, for an evaluation that is to include all possible combinations among n interventions, 2^n groups would be needed. In addition, to be able to distinguish differences in outcomes among the different groups, each group must contain a sufficient number of units of observation to ensure sufficient statistical power. In practice, detecting differences between different intervention arms may require larger samples than when comparing a treatment to a pure comparison. If the two treatment arms are successful in causing changes in the desired outcomes, larger samples will be required to detect the potentially minor differences between the two groups.[3]

Key Concept

For an evaluation to evaluate the impact of all possible combinations among n different interventions, you will need a total of 2^n treatment and comparison groups.

Finally, crossover designs can also be put in place in evaluation designs that combine various evaluation methods. The operational rules that guide the assignment of each treatment will determine which combination of methods has to be used. For instance, it may be that the first treatment is allocated based on an eligibility score, but the second one is allocated in a randomized fashion. In that case, the design can use a regression discontinuity design for the first intervention and a randomized assignment method for the second intervention.

Additional Resources

- For accompanying material to the book and hyperlinks to additional resources, please see the Impact Evaluation in Practice website (http://www.worldbank.org/ieinpractice).
- For more information on impact evaluation design with multiple treatment options, see Banerjee, Abhijit, and Esther Duflo. 2009. "The Experimental Approach to Development Economics." *Annual Review of Economics* 1: 151–78.

Notes

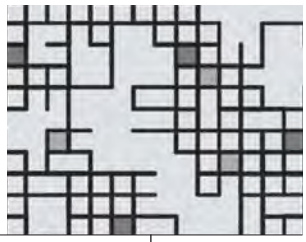
1. See Banerjee and Duflo (2009) for a longer discussion.
2. Note that in practice, it is possible to combine the three separate lotteries into one and still achieve the same result.
3. Testing the impact of multiple interventions also has a more subtle implication: as we increase the number of interventions or levels of treatment that we test against one another, we increase the likelihood that we will find an impact in at least one of the tests, even if there is no impact. In other words, we are more likely to find a false positive result. To prevent this, statistical tests must be adjusted to account for multiple hypothesis testing. False positives are also referred to as type II errors. See chapter 15 for more information on type II errors and references on multiple hypothesis testing.

References

Banerjee, Abhijit, and Esther Duflo. 2009. "The Experimental Approach to Development Economics." *Annual Review of Economics* 1: 151–78.

Olken, Benjamin. 2007. "Monitoring Corruption: Evidence from a Field Experiment in Indonesia." *Journal of Political Economy* 115 (2): 200–249.

Pop-Eleches, Cristian, Harsha Thirumurthy, James Habyarimana, Joshua Zivin, Markus Goldstein, Damien de Walque, Leslie MacKeen, Jessica Haberer, Sylvester Kimaiyo, John Sidle, Duncan Ngare, and David Bangsberg. 2011. "Mobile Phone Technologies Improve Adherence to Antiretroviral Treatment in a Resource-Limited Setting: A Randomized Controlled Trial of Text Message Reminders." *AIDS* 25 (6): 825–34.



Part 3

HOW TO IMPLEMENT AN IMPACT EVALUATION

Part 3 of this book focuses on how to implement an impact evaluation: how to select an impact evaluation method compatible with a program's operational rules; how to manage an impact evaluation, including ensuring a strong partnership between the research and policy teams and managing the time and budget for an evaluation; how to ensure that an evaluation is both ethical and credible, following principles for working with human subjects and open science; and how to use impact evaluation to inform policy.

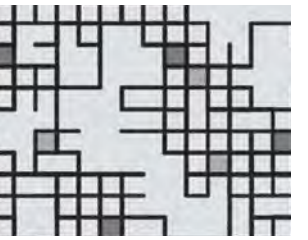
Chapter 11 outlines how to use the rules of program operation—namely, a program's available resources, criteria for selecting beneficiaries, and timing for implementation—as the basis for selecting an impact evaluation method. A simple framework is set out to determine which of the impact evaluation

methodologies presented in part 2 is most suitable for a given program, depending on its operational rules. The chapter further discusses how the preferred method is the one that requires the weakest assumptions and has the fewest data requirements within the context of the operational rules.

Chapter 12 discusses the relationship between the research and policy teams and their respective roles. It reviews the distinction between independence and unbiasedness, and highlights areas that may prove to be sensitive in carrying out an impact evaluation. The chapter provides guidance on how to manage stakeholders' expectations and highlights some of the common risks involved in conducting impact evaluations, as well as suggestions on how to manage those risks. The chapter concludes with an overview of how to manage impact evaluation activities, including setting up the evaluation team, timing the evaluation, budgeting, and fundraising.

Chapter 13 provides an overview of the ethics and science of impact evaluation, including the importance of not denying benefits to eligible beneficiaries for the sake of the evaluation; how to apply core principles of ethical research involving human subjects; the role of institutional review boards that approve and monitor research involving human subjects; and the importance of practicing open science, including registering evaluations and making data publically available for further research and for replicating results.

Chapter 14 provides insights into how to use impact evaluations to inform policy, including tips on how to make the results relevant, a discussion of the kinds of products that impact evaluations can and should deliver, and guidance on how to produce and disseminate findings to maximize policy impact.



CHAPTER 11

Choosing an Impact Evaluation Method

Determining Which Method to Use for a Given Program

The key to identifying the causal impact of a program is finding a valid comparison group to estimate the counterfactual and answer the policy question of interest. In part 2, we discussed a number of methods, including randomized assignment, instrumental variables, regression discontinuity design, difference-in-differences, and matching. In this chapter, we consider the question of which method to choose for a given program that you would like to evaluate.

First, we show that the program's operational rules provide clear guidance on how to find comparison groups, and thus on which method is most appropriate for your policy context. An overarching principle is that, if the operational rules of a program are well defined, then they can help to determine which method is best suited to evaluate that particular program.

Second, the methods introduced in part 2 have different data requirements and rely on different underlying assumptions. Some methods require stronger assumptions than others to precisely estimate the changes in outcomes caused by the intervention. In general, we prefer the method that

requires the weakest assumptions and has the fewest data requirements within the context of the operational rules.

Finally, we discuss how to choose the unit of intervention. For example, will the program be assigned at the individual level or at a higher level, such as communities or districts? In general, we prefer choosing the smallest unit of intervention feasible within operational constraints.

How a Program's Rules of Operation Can Help Choose an Impact Evaluation Method

Key Concept

The operational rules of a program determine which impact evaluation method is best suited to evaluate that program—not vice versa.

One of the main messages of this book is that we can use a program's operational rules to find valid comparison groups, to the extent that the program operational rules are well defined. In fact, the rules of program operation provide a guide to which method is best suited to evaluate that particular program. It is the program rules of operations that can and should drive the evaluation method—not vice versa. The evaluation should not drastically change key elements of well-defined program assignment rules for the sake of a cleaner evaluation design.

The operational rules most relevant for the evaluation design are those that identify who is eligible for the program and how they are selected for participation. Comparison groups come from those that are eligible but cannot be incorporated at a given moment (for example, when there are resource constraints and excess demand exists), or those near an eligibility threshold for participation in the program.

Principles for Well-Defined Program Assignment Rules

Key Concept

When designing prospective impact evaluations, we can almost always find valid comparison groups if the operational rules for selecting beneficiaries are equitable, transparent, and accountable.

When designing prospective impact evaluations, we can almost always find valid comparison groups if the operational rules for selecting beneficiaries are equitable, transparent, and accountable:

- *Equitable* program assignment rules rank or prioritize eligibility based on a commonly agreed indicator of need, or stipulate that everyone is offered program benefits or at least has an equal chance of being offered benefits.
- *Transparent* program assignment rules are made public, so that outside parties can implicitly agree to them and can monitor that they are actually followed. Transparent rules should be quantitative and easily observable.
- *Accountable* rules are the responsibility of program officials, and their implementation is the basis of the job performance or rewards of those officials.

The operational rules of eligibility are transparent and accountable when programs use quantifiable criteria that can be verified by outside parties and when they make those criteria public. Equity, transparency, and accountability ensure that eligibility criteria are quantitatively verifiable and are actually implemented as designed. As such, these principles of good governance improve the likelihood that the program will actually benefit the target population, and they are also the key to a successful evaluation. If the rules are not quantifiable and verifiable, then the evaluation team will have difficulty making sure that assignment to treatment and comparison groups happens as designed or, at minimum, documenting how it actually happened. If members of the evaluation team cannot actually verify assignment, then they cannot correctly analyze the data to calculate impacts. Understanding the program assignment rules is critical to selecting a proper impact evaluation method.

When the operational rules violate any of these three principles of good governance, we face challenges both in creating a well-designed program and in conducting the evaluation. It is difficult to find valid comparison groups if the rules that determine beneficiaries' eligibility and selection are not equitable, transparent, and accountable. In this case, the design of an impact evaluation may require clarifications and adjustments in the way the program operates. If the rules are well defined, however, the impact evaluation method can be chosen based on the existing program assignment rules, as we now discuss in more detail.

Key Operational Rules

Rules of operation typically govern what the program benefits are, how those benefits are financed and distributed, and how the program selects beneficiaries. The rules governing programs and the selection of beneficiaries are key to finding valid comparison groups. The rules governing beneficiary selection cover eligibility, allocation rules in the case of limited resources, and the phasing in of beneficiaries over time. More specifically, the key rules that generate a road map to find comparison groups answer three fundamental operational questions relating to a program's available resources, eligibility criteria, and timing for implementation:

1. *Available resources.* Does the program have sufficient resources to achieve scale and reach full coverage of all eligible beneficiaries? Governments and nongovernmental organizations do not always have sufficient resources to provide program services to everyone who is eligible and applies for benefits. In that case, the government must decide which of the eligible applicants will receive program benefits and which will

not be included. Many times, programs are limited to specific geographic regions, or to a limited number of communities, even though there may be eligible beneficiaries in other regions or communities.

2. *Eligibility criteria*. Who is eligible for program benefits? Is the program assignment based on an eligibility cutoff, or is it available to everyone? Public school and primary health care are usually offered universally. Many programs use operational eligibility rules that rely on a continuous ranking with a cutoff point. For example, pension programs set an age limit above which elderly individuals become eligible. Cash transfer programs often rank households based on their estimated poverty status, and households below a predetermined cutoff are deemed eligible.

3. *Timing for implementation*. Are potential beneficiaries enrolled in the program all at once, or in phases over time? Often, administrative and resource constraints prevent governments and nongovernmental organizations from immediately providing benefits to the entire eligible population. They must roll out their programs over time, and thus must decide who gets the benefits first and who is incorporated later. A common approach is to phase in a program geographically, over time, incorporating all eligible beneficiaries in one village or region before moving to the next.

Deriving Comparison Groups from Operational Rules

When designing prospective impact evaluations, answering the three operational questions largely determines the impact evaluation method that is suitable for a given program. Table 11.1 maps the possible comparison groups to specific program operational rules and the three fundamental operational questions related to available resources, eligibility rules, and timing for implementation. The columns are split as to whether or not the program has sufficient resources to cover all potentially eligible beneficiaries eventually (*available resources*), and are further subdivided into programs that have a continuous eligibility ranking and cutoff and those that do not (*eligibility criteria*). The rows are divided into phased versus immediate implementation of the program (*timing for implementation*). Each cell lists the potential sources of valid comparison groups, along with the related chapter in which they were discussed in part 2. Each cell is labeled with an index: the initial letter indicates the row in the table (A, B), and the number that follows indicates the column (1–4). For example, cell A1 refers to the cell in the first row and first column of the table. For instance, cell A1 identifies the evaluation methods that are most adequate for programs that have limited resources, have eligibility criteria, and are phased in over time.

Table 11.1 Relationship between a Program's Operational Rules and Impact Evaluation Methods

Timing of Implementation		Excess demand for program (limited resources)		No excess demand for program (fully resourced)	
	Eligibility criteria	(1) Continuous eligibility ranking and cutoff	(2) No continuous eligibility ranking and cutoff	(3) Continuous eligibility ranking and cutoff	(4) No continuous eligibility ranking and cutoff
	(A) Phased implementation over time	Cell A1 Randomized assignment (chapter 4) RDD (chapter 6)	Cell A2 Randomized assignment (chapter 4) Instrumental variables (randomized promotion) (chapter 5) DD (chapter 7) DD with matching (chapter 8)	Cell A3 Randomized assignment to phases (chapter 4) RDD (chapter 6)	Cell A4 Randomized assignment to phases (chapter 4) Instrumental variables (randomized promotion to early take-up) (chapter 5) DD (chapter 7) DD with matching (chapter 8)
	(B) Immediate implementation	Cell B1 Randomized assignment (chapter 4) RDD (chapter 6)	Cell B2 Randomized assignment (chapter 4) Instrumental variables (randomized promotion) (chapter 5) DD (chapter 7) DD with matching (chapter 8)	Cell B3 RDD (chapter 6)	Cell B4 If less than full take-up: Instrumental variables (randomized promotion) (chapter 5) DD (chapter 7) DD with matching (chapter 8)

Note: DD = difference-in-differences; RDD = regression discontinuity design.

Most programs need to be phased in over time because of either financing constraints or logistical and administrative limitations. This group or category covers the first row of the chart (cells A1, A2, A3, and A4). In this case, the equitable, transparent, and accountable operational rule is to give every eligible unit an equal chance of getting the program first, second, third, and so on, implying randomized rollout of the program over time.

In the cases in which resources are limited—that is, in which there will never be enough resources to achieve full scale-up (cells A1 and A2, and B1 and B2)—excess demand for those resources may emerge very quickly. Then a lottery to decide who gets into the program may be a viable approach to assign benefits among equally eligible units. In this case, each eligible unit gets an equal chance to benefit from the program. A lottery is an example

of an equitable, transparent, and accountable operational rule to allocate program benefits among eligible units.

Another class of programs comprises those that are phased in over time and for which administrators can rank the potential beneficiaries by need (cells A1 and A3). If the criteria used to prioritize the beneficiaries are quantitative and available and have a cutoff for eligibility, the program can use a regression discontinuity design.

The other broad category consists of programs that have the administrative capability to be implemented immediately: that is, the cells in the bottom row of the chart. When the program has limited resources and is not able to rank beneficiaries (cell B2), then randomized assignment based on excess demand could be used. If the program has sufficient resources to achieve scale and no eligibility criteria (cell B4), then the only solution is to use instrumental variables (randomized promotion), under the assumption of less than full take-up of the program. If the program can rank beneficiaries and relies on eligibility criteria, regression discontinuity design can be used.

Prioritizing Beneficiaries

All three key operational questions relate to the critical issue of how beneficiaries are selected, which is crucial to find valid comparison groups. Comparison groups are sometimes found among the ineligible populations, and more frequently among the populations that are eligible but are incorporated into the program later. How beneficiaries are prioritized depends in part on the objectives of the program. Is it a pension program for the elderly, a poverty alleviation program targeted to the poor, or an immunization program available to everyone?

To prioritize beneficiaries based on need, the program must find an indicator that is both quantifiable and verifiable. In practice, feasibility of prioritization depends largely on the ability of the government to measure and rank need. If the government can accurately rank beneficiaries based on relative need, it may feel ethically obligated to roll out the program in order of need. However, ranking based on need requires not only a quantifiable measure, but also the ability and resources to measure that indicator for each unit that could participate in the program.

Some programs use selection criteria that could in principle be used to rank relative need and determine eligibility. For example, many programs seek to reach poor individuals. However, accurate poverty indicators that reliably rank households are often hard to measure and costly to collect. Collecting income or consumption data on all potential beneficiaries to rank them by poverty level is a complex and costly process, which would also be hard to verify. Instead, many programs use some sort of proxy means test

to estimate poverty levels. These are indexes of simple observable measures such as assets and sociodemographic characteristics (Grosh and others 2008). Proxy means tests can help determine reasonably well whether a household is above or below some gross cutoff, but they may be less precise in providing a detailed ranking of socioeconomic status or need.

Rather than confront the cost and complexity of ranking potential individual beneficiaries, many programs choose to rank at a higher level of aggregation, such as at the community level. Determining program assignment at an aggregate level has obvious operational benefits, but it is often difficult to find indicators to achieve a ranking of needs at a more aggregate level.

In cases when a program cannot reliably assign benefits based on need, because a quantifiable and verifiable ranking indicator either is not available or is too costly and prone to error, other criteria need to be used to decide how to sequence program rollout. One criterion that is consistent with good governance is equity. An equitable rule would be to give everyone who is eligible an equal chance of going first, and as such to randomly assign a place in the sequence to potential beneficiaries. In practice, given the challenges in ranking need, randomized assignment of program benefits is a commonly used program assignment rule. It is a fair and equitable allocation rule. It also produces a randomized evaluation design that can provide good internal validity if implemented well, and it can rely on weaker assumptions compared with the other methods, as discussed in the next section.

A Comparison of Impact Evaluation Methods

After assessing which impact evaluation method is suitable for specific program operational rules, the evaluation team can choose the method that has the weakest assumption and fewest data requirements. Table 11.2 provides a comparison of the alternative impact evaluation methods in terms of the data requirements to implement them, and the underlying assumptions necessary to interpret their results as causal impacts of the intervention. Each row represents a separate method. The first two columns describe the methods and the units that are in the comparison group. The last two columns report the assumptions needed to interpret the results as causal and the data needed to implement the methods.

All methods require assumptions; that is, to be able to interpret results as causal, we must believe facts to be true that we cannot always fully verify empirically. In particular, for each method, one key assumption is that the mean of the comparison group on which the method relies is a valid

Table 11.2 Comparing Impact Evaluation Methods

Methodology	Description	Who is in the comparison group?	Key assumption	Required data
Randomized assignment	Eligible units are randomly assigned to a treatment or comparison group. Each eligible unit has an equal chance of being selected. Tends to generate internally valid impact estimates under the weakest assumptions.	Eligible units that are randomly assigned to the comparison group.	Randomization effectively produces two groups that are statistically identical with respect to observed and unobserved characteristics (at baseline and through endline).	Follow-up outcome data for treatment and comparison groups; baseline outcomes and other characteristics for treatment and comparison groups to check balance.
Instrumental variable (particularly randomized promotion)	A randomized instrument (such as a promotion campaign) induces changes in participation in the program being evaluated. The method uses the change in outcomes induced by the change in participation rates to estimate program impacts.	"Complier" units whose participation in the program is affected by the instrument (they would participate if exposed to the instrument, but would not participate if not exposed to the instrument).	The instrument affects participation in the program but does not directly affect outcomes (that is, the instrument affects outcomes only by changing the probability of participating in the program).	Follow-up outcome data for all units; data on effective participation in the program; data on baseline outcomes and other characteristics.
Regression discontinuity design	Units are ranked based on specific quantitative and continuous criteria, such as a poverty index. There is a cutoff that determines whether or not a unit is eligible to participate in a program. Outcomes for participants on one side of the cutoff are compared with outcomes for nonparticipants on the other side of the cutoff.	Units that are close to the cutoff but are ineligible to receive the program.	To identify unbiased program impacts for the population close to the cutoff, units that are immediately below and immediately above the cutoff are statistically identical. To identify unbiased program impacts for the whole population, the population close to the cutoff needs to be representative of the whole population.	Follow-up outcome data; ranking index and eligibility cutoff; data on baseline outcomes and other characteristics.

(continued)

Table 11.2 *(continued)*

Methodology	Description	Who is in the comparison group?	Key assumption	Required data
Difference-in-differences	The change in outcome over time in a group of nonparticipants is used to estimate what would have been the change of outcomes for a group of participants in the absence of a program.	Units that did not participate in the program (for any reason), and for which data were collected before and after the program.	If the program did not exist, outcomes for the groups of participants and nonparticipants would have grown in parallel over time.	Baseline and follow-up data on outcomes and other characteristics for both participants and nonparticipants.
Matching (particularly propensity score matching)	For each program participant, the method looks for the "most similar" unit in the group of nonparticipant (the closest match based on observed characteristics).	For each participant, the nonparticipant unit that is predicted to have the same likelihood to have participated in the program based on observed characteristics.	There is no characteristic that affects program participation beyond the observed characteristics used for matching.	Follow-up outcome data for participants and nonparticipants; data on effective participation in the program; baseline characteristics to perform matching.

Source: Adapted from the Abdul Latif Jameel Poverty Action Lab (J-PAL) website.

estimate of the counterfactual. In each of the chapters on methods in part 2, we discussed some considerations of how to test whether a method is valid in a particular context. Some methods rely on stronger assumptions than others.

Key Concept

The preferred impact evaluation method is the one that best fits the operational context, requires the weakest assumptions, and has the fewest data requirements.

All other things equal, the method that best fits the operational context, and that requires the weakest assumptions and the least data, is the preferred method. These criteria explain why researchers settle on randomized assignment as the gold standard, and why it is often the preferred method. Randomized assignment fits many operational contexts, and it tends to generate internally valid impact estimates under the weakest assumptions. When properly implemented, it generates comparability between the treatment and comparison groups in observed and unobserved characteristics. In addition, randomized assignment tends to require smaller samples than the samples needed to implement quasi-experimental methods (see discussion in chapter 15). Because randomized assignment is fairly intuitive, the method also makes it straightforward to communicate results to policy makers.

Quasi-experimental methods may be more suitable in some operational contexts, but they require more assumptions in order for the comparison group to provide a valid estimate of the counterfactual. For example, difference-in-differences relies on the assumption that changes in outcomes in the comparison group provide a valid estimate of the counterfactual changes in outcomes for the treatment group. This assumption that the outcomes in the treatment and comparison groups grow in parallel over time is not always possible to test without multiple waves of data before the intervention. Regression discontinuity relies on comparability of units just below and just above the eligibility threshold. Matching has the strongest assumptions of all methods, as it essentially assumes away any unobserved characteristics between program participants and nonparticipants. Overall, the stronger the assumptions, the higher the risk for them not to hold in practice.

A Backup Plan for Your Evaluation

Sometimes things do not go exactly as planned, even with the best impact evaluation design and the best intentions. In one job training program, for example, the implementation agency planned to randomly select participants from the pool of applicants, based on the expected oversubscription to the program. Because unemployment among the target population was high, it was anticipated that the pool of applicants for the job training program would be much larger than the number of places available. Unfortunately, advertisement for the program was not as effective as expected, and in the end, the number of applicants was just below the

number of training slots available. Without oversubscription from which to draw a comparison group, and with no backup plan in place, the initial attempt to evaluate the program had to be dropped entirely. This kind of situation is common, as are unanticipated changes in the operational or political context of a program. Therefore, it is useful to have a backup plan in case the first choice of methodology does not work out.

Planning to use several impact evaluation methods is also good practice from a methodological point of view. If you have doubts about whether one of your methods may have remaining bias, you will be able to check the results against the other method. When a program is implemented in a randomized rollout, the comparison group will eventually be incorporated into the program. That limits the time during which the comparison group is available for the evaluation. If, however, in addition to the randomized assignment design, a randomized promotion design is also implemented, then a comparison group will be available for the entire duration of the program. Before the final group of the rollout is incorporated, two alternative comparison groups will exist (from the randomized assignment and the randomized promotion), though in the longer term only the randomized promotion comparison group will remain.

Finding the Smallest Feasible Unit of Intervention

In general, the rules of operation also determine the level at which an intervention is assigned, which relates to the way the program is implemented. For example, if a health program is implemented at the district level, then all villages in the district would either receive the program (as a group) or not receive it. Some programs can be efficiently implemented at the individual or household level, whereas others need to be implemented at a community or higher administrative level. Even if a program can be assigned and implemented at the individual level, the evaluation research team may prefer a higher level of aggregation in order to mitigate potential spillovers, that is, indirect effects from participating to nonparticipating units (see discussion in chapter 9).

Implementing an intervention at a higher level can be problematic for the evaluation for two main reasons. First, evaluations of interventions assigned and implemented at higher levels, such as the community or administrative district, require larger sample sizes and will be more costly compared with evaluations of interventions at a lower level, such as at the individual or household level. The level of intervention is important because

it defines the unit of assignment to the treatment and comparison groups, and that has implications for the size of the evaluation sample and its cost. For interventions implemented at higher levels, a larger sample is needed to be able to detect the program's true impact. The intuition behind this will be discussed in chapter 15, which reviews how to determine the sample size required for an evaluation and discusses how implementation at higher levels creates clusters that increase the required sample size.

Second, at higher levels of intervention, it is harder to find a sufficient number of units to include in the evaluation. Yet randomized assignment only generates comparable treatment and comparison groups if it is performed among a sufficient number of units. For example, if the level of aggregation is at the province level and the country has only six provinces, then randomization is unlikely to achieve balance between the treatment and comparison groups. In this case, say that the evaluation design allocates three states to the treatment group and three to the comparison group. It is very unlikely that the states in the treatment group would be similar to the comparison group, even if the number of households within each state is large. This is because the key to balancing the treatment and comparison groups is the number of units assigned to the treatment and comparison groups, not the number of individuals or households in the sample. Therefore, performing randomized assignment at high levels of implementation creates risks for internal validity if the number of units is not sufficient.

To avoid the risks associated with implementing an intervention at a high geographical or administrative level, the evaluation team and program managers need to work together to find the smallest unit of intervention that is operationally feasible. Various factors determine the smallest feasible unit of intervention:

- Economies of scale and administrative complexity in the delivery of the program
- Administrative ability to assign benefits at the individual or household level
- Potential concerns about possible tensions
- Potential concerns about spillovers and contamination of the comparison group.

The smallest feasible unit of intervention typically depends on economies of scale and the administrative complexity of delivering the program. For example, a health insurance program may require a local office for

beneficiaries to submit claims and to pay providers. The fixed costs of the office need to be spread over a large number of beneficiaries, so it might be inefficient to roll out the program at the individual level and more efficient to do so at the community level. However, in situations with new and untested types of interventions, it may be worth absorbing short-run inefficiencies and rolling out the program within administrative districts, so as to better ensure credibility of the evaluation and lower the costs of data collection.

Some program managers argue that locally administered programs, such as health insurance programs, do not have the administrative capabilities to implement programs at the individual level. They worry that it would be a burden to set up systems to deliver different benefits to different beneficiaries within local administrative units, and that it may be challenging to guarantee that the assignment of treatment and comparison groups will be implemented as designed. The latter issue is a serious threat for an impact evaluation, as program managers may not be able to implement the program consistently with an evaluation design. In this case, implementation at a higher level or simplification of the impact evaluation design may be necessary.

Sometimes governments prefer to implement programs at more aggregate levels, such as the community, because they worry about potential tensions when members of the comparison group observe neighbors in the treatment group getting benefits. Many programs have been successfully implemented at the individual or household level within communities without generating tensions, in particular when benefits have been assigned in an equitable, transparent, and accountable way. Still, the risk that tensions may arise needs to be considered in the context of a specific impact evaluation.

Finally, when a program is assigned and implemented at a very low level, such as the household or individual level, contamination of the comparison group may compromise the internal validity of the evaluation. For example, say that you are evaluating the effect of providing tap water on households' health. If you install the taps for a household but not for its neighbor, the treatment household may well share the use of the tap with a comparison neighbor; the neighboring household then would not be a true comparison, since it would benefit from a spillover effect.

Box 11.1 illustrates the implications of the choice of implementation level of intervention in the context of cash transfer programs. In practice, program managers therefore need to choose the smallest feasible unit of intervention that (1) allows a large enough number of units for the evaluation, (2) mitigates the risks to internal validity, and (3) fits the operational context.

Box 11.1: Cash Transfer Programs and the Minimum Level of Intervention

The majority of conditional cash transfers use communities as the unit or level of intervention, for administrative and program design reasons, as well as out of concern about spillovers and potential tensions in the community if treatment were to be assigned at a lower level.

For example, the evaluation of Mexico's conditional cash transfer program, Progresa/Oportunidades, relied on the rollout of the program at the community level in rural areas to randomly assign communities to the treatment and comparison groups. All eligible households in the treatment communities were offered the opportunity to enroll in the program in spring 1998, and all eligible households in the comparison communities were offered the same opportunity 18 months later, in winter 1999. However, the evaluation team found substantial correlation in outcomes between households within communities. Therefore, to generate sufficient statistical power for the evaluation, they needed more households in the sample than would have been needed if they had been able to assign individual households to the treatment and comparison groups. The impossibility of implementing the program at the household level therefore led to larger sample size requirements and increased the cost of the evaluation. Similar constraints apply to many programs in the human development sector.

Sources: Behrman and Hoddinott 2001; Skoufias and McClafferty 2001.

Additional Resources

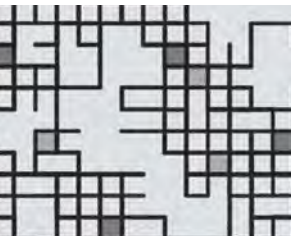
- For accompanying material to the book and hyperlinks to additional resources, please see the Impact Evaluation in Practice website (http://www.worldbank.org/ieinpractice).

References

Behrman, Jere R., and John Hoddinott. 2001. "An Evaluation of the Impact of PROGRESA on Preschool Child Height." Discussion Paper No. 104, Food Consumption and Nutrition Division, International Food Policy Research Institute, Washington, DC.

Grosh, M. E., C. Del Ninno, E. Tesliuc, and A. Ouerghi. 2008. *For Protection and Promotion: The Design and Implementation of Effective Safety Nets*. Washington, DC: World Bank.

Skoufias, Emmanuel, and Bonnie McClafferty. 2001. "Is *Progresa* Working? Summary of the Results of an Evaluation by IFPRI." International Food Policy Research Institute, Washington, DC.



CHAPTER 12

Managing an Impact Evaluation

Managing an Evaluation's Team, Time, and Budget

An evaluation is a partnership between a policy team and a research team. Each group depends on the other for the success of the evaluation. Together, they constitute the evaluation team. The partnership is based on an understanding of the respective roles and responsibilities of the two teams, a joint commitment to the evaluation, and a recognition of what motivates people's work on the evaluation. An effective partnership is critical to ensuring the technical credibility and policy impact of an evaluation.

This chapter outlines elements of an effective partnership, including the roles and responsibilities of each team. It explores how the partnership works at different stages of the evaluation process and reviews alternative models for collaboration. The chapter also addresses practical questions of timing and budgeting.

Roles and Responsibilities of the Research and Policy Teams

The Research Team: Research Function and Data Function

The research team is responsible for the technical quality and scientific integrity of the evaluation work. Its responsibilities encompass research design, data quality, and analysis. Research teams typically comprise the following people:

- The *principal investigator* works with policy makers and program implementers to establish the key objectives, policy questions, indicators, and information needs of the evaluation (often using a theory of change as depicted by a results chain); determine the impact evaluation methodology; develop the evaluation plan; identify the research team; register the impact evaluation; obtain approvals from the Institutional Review Board (IRB); prepare an evaluation plan, including a more detailed preanalysis plan; lead the analysis of results; and engage with the policy team to disseminate results. The principal investigator needs to be able to work effectively with the full evaluation team, including the organization in charge of data collection, other members of the research team, and policy makers or program implementers who use the data and the results of the evaluation. A number of *researchers* may work with the principal investigator or as co-principal investigators to lead or support specific analytical work on elements, such as sampling, qualitative assessment, or cost-effectiveness analysis.
- An *evaluation manager or field coordinator* works directly with the principal investigator on the day-to-day implementation of the evaluation. This includes working with program implementers and policy makers on the policy team and overseeing fieldwork when primary data are being collected. This person is particularly important in cases where the principal investigator is not based locally, where a prospective evaluation is being applied that needs to be closely coordinated with program implementation, or where primary data are being collected.
- A *sampling expert* guides work on power calculations and sampling. For the type of quantitative impact evaluation covered in this book, the sampling expert should be able to carry out power calculations to determine the appropriate sample sizes for the indicators established; select the sample; review the results of the actual sample versus the designed sample; and provide advice on implications for the analysis in line with the preanalysis plan. The principal investigator often performs these functions directly or together with the sampling expert.

- A *data collection team* is responsible for developing data collection instruments and accompanying manuals and codebooks; collecting, digitizing, and cleaning the data; and delivering a clean and documented data set, when primary data collection is required. Chapter 16 discusses data sources and various aspects of data collection.

The Policy Team: Policy Function and Program Management Function

The policy team consists of policy makers and program implementers:

- *Policy makers* set the research agenda, identify the core study question to be addressed, ensure adequate resources are available for the work, and apply the results to policy. At the outset of the evaluation, they need to clearly articulate the objectives of both the program and the evaluation, as well as the theory of change and the main indicators of interest, including the minimum policy-relevant effect size for the outcome indicators of interest, as outlined in chapter 2. The policy team has the knowledge of the policy dialogue and contacts with key stakeholders to ensure that the evaluation is designed to be as policy-relevant as possible, and to ensure that the appropriate stakeholders and decision makers are engaged at key points in the evaluation process.
- *Program implementers* work hand in hand with the research team to align the evaluation design and program implementation. This includes verifying that the evaluation design is based on accurate information about the program's operation, and committing to implement the program as planned, in the case of prospective evaluations. Program implementers on the policy team also typically manage the evaluation budget and are often engaged in helping the research team supervise fieldwork for data collection.

Who Cares about the Evaluation and Why?

Key Concept

An effective partnership between the policy team and the research team is critical to ensuring the technical credibility and policy impact of an evaluation.

From the perspective of the policy team, the primary interest is usually to know whether or not the program or reform is effective, and at what cost the results were achieved, thereby allowing the team to make policy decisions on the basis of the evidence produced. The local program implementers will be interested in ensuring that their efforts are valued and that they get credit and visibility for their work, which often reaches beyond the boundaries of their day-to-day responsibilities. A good way to value these contributions is to ensure that local teams are actively engaged in the broader range of evaluation activities. This can be done by holding joint workshops, as well as by

issuing joint publications, ensuring training and capacity building, and engaging local researchers who are well placed to contribute substantively and can serve as an important conduit between the research and policy teams.

Evaluations have value in terms of a public good when they inform a question of interest beyond the immediate interest of the policy team. This aspect is often of primary interest to researchers exploring fundamental questions pertaining to a theory of change. For example, results concerning how people behave under certain circumstances or how transmission channels function, allowing impacts to be achieved, can allow more general lessons to be drawn and applied in different settings. Impact evaluations are rapidly contributing to a global evidence base on the performance of a range of program and policy reforms, constituting repositories of knowledge highly relevant to program and policy design. Donors and policy institutes are often interested in this broader public good value and are increasingly providing financial support to conduct evaluations that contribute to this evidence base.

Researchers will also be very committed to the use of a robust, defensible evaluation methodology and will want to ensure that they are engaged in the design of the impact evaluation, in the analysis of the data, and in the generation of primary research that meets scientific standards for publication in academic journals. Interdisciplinary research teams have an added challenge of ensuring that there is a common understanding among team members. Different disciplines, such as medicine and economics, may have different approaches to registering trials, engaging subjects, reporting results, or disseminating results, for example. These different expectations are best clarified and understood at the outset of an evaluation. Regardless of different protocols, research teams are expected to follow generally accepted scientific norms and ethical principles, discussed in chapter 13.

The different interests of the policy team and the research team can create tensions that need to be understood and managed. Researchers tend to value technical rigor in the evaluation design over the operational feasibility of program implementation. The teams may also be interested in somewhat different evaluation questions. Finally, neither team may be interested in publishing nuanced or negative results, as this may reflect poorly on the program performance for the policy team and may be of less academic interest to the research team. The policy team may also be interested in being selective about which results are released, whereas the research team will value highly the ability to publish the full range of results.

For the evaluation team as a whole, fostering a culture of transparency and respect for evidence is critical. Policy makers and program managers should be rewarded for their commitments to evidence-based policy making. Even when results are not favorable, these actors should be credited for

having championed transparency. Likewise, the research team should be encouraged to report on and publish results, regardless of the findings.

The Research and Policy Team Partnership during the Evaluation

The technical quality and policy impact of the evaluation depend on an active partnership between the research team and the policy team at each stage in the evaluation: design, implementation, analysis, and dissemination. Box 12.1 summarizes some guiding principles.

Design stage. First, the policy makers need to clearly structure and convey the core research questions, the accompanying theory of change, and the core indicators of interest, and ensure that the research team has a good understanding of and respect for these elements. To ensure policy relevance, the policy team also needs to take the lead in structuring an engagement strategy that will ensure that the necessary stakeholders are consulted and informed about the design, implementation, and results of the evaluation. For their part, researchers need to clarify for the policy team the necessary conditions for good impact evaluations. In the case of prospective evaluations, this will first involve verifying with the program implementers and policy makers in the policy team that program operations are well enough established to ensure that the program being evaluated will not change a great deal during the evaluation—and thus will not render the results irrelevant for policy purposes. The "sweet spot" for conducting an impact evaluation is often the point at which the program has been field tested enough to establish that it is operating in the manner intended—which can be informed

Box 12.1: Guiding Principles for Engagement between the Policy and Evaluation Teams

- Engage early to maximize evaluation design options and ensure an effective partnership between the policy and evaluation teams.
- Have a clear impact evaluation plan at the outset.
- Understand roles, responsibilities, and motivations of the various stakeholders and give them a stake in the evaluation.
- Stay engaged throughout the evaluation to ensure the proper alignment between the evaluation and the intervention being evaluated.
- Acknowledge and manage risks and benefits, being clear about what impact evaluations can and cannot do.
- Value transparency and ensure objectivity and be prepared to respect the results, good or bad.

by a good process evaluation—but has not been expanded, thereby leaving more options to construct appropriate counterfactuals.

Second, the research team needs to clearly understand the program's rules of operation: namely, its available resources, eligibility criteria for selecting beneficiaries, and timing for implementation. The policy team should clearly convey these three rules of operation to the research team, as these are fundamental to informing the methodological options available in the evaluation, as detailed in chapter 11.

Third, the research team should prepare an impact evaluation plan that contains both operational and research aspects, and share this with policy makers to ensure that the evaluation is focused on the questions of interest; that elements of collaboration with the policy team are outlined; and that the evaluation team is clear and straightforward about the questions being asked and the nature and timing of the results (see box 12.2). It is also useful to consider risks and proposed mitigation strategies. Finally, the research team should obtain ethical approval from an institutional review board and register the evaluation in a trial registry (see chapter 13).

This dialogue during the design stage should result in a clear, shared commitment to an evaluation plan, with realistic expectations and mutually agreed upon responsibilities for members of the policy and research teams. This dialogue provides an opportunity for the research team to clarify both the value of an impact evaluation—notably the establishment of causality and the generalizability of the findings—and its limitations, such as not providing insights into why certain results are obtained, trade-offs with sample sizes and power calculations, or the time involved in generating certain results. This dialogue also provides the opportunity for the policy team to specify priority questions and to ensure that the evaluation is well aligned with policy questions of interest.

Implementation stage. The policy and research teams need to work together to ensure that implementation proceeds smoothly and to troubleshoot. For example, in a randomized controlled trial, the teams need to agree on the best way to randomize in practice. In addition, during this stage, coordination is especially important to ensure fidelity between the evaluation design and program implementation.

Analysis stage. The analysis that is carried out should correspond to what is outlined in the evaluation plan and in the more detailed preanalysis plan. The research team should provide and discuss results with the policy team at key junctures. As early as the baseline, this should include a review of the quality of the data collected and adherence to the evaluation plan. This will help ensure that the evaluation plan envisioned in the design stage remains feasible and allow any necessary adjustments to be made. This is also an excellent opportunity to review which products

Box 12.2: General Outline of an Impact Evaluation Plan

1. Introduction
2. Description of the intervention
3. Objectives of the evaluation
   - 3.1 Hypotheses, theory of change, results chain
   - 3.2 Policy questions
   - 3.3 Key outcome indicators
   - 3.4 Risks
4. Evaluation design
5. Sampling and data
   - 5.1 Sampling strategy
   - 5.2 Power calculations
6. Preanalysis plan overview
7. Data collection plan
   - 7.1 Baseline survey
   - 7.2 Follow-up survey(s)
8. Products to be delivered
   - 8.1 Baseline report
   - 8.2 Impact evaluation report
   - 8.3 Policy brief
   - 8.4 Fully documented data sets, design and analysis protocols
9. Dissemination plan
10. Ethical protocols on protection of human subjects
    - 10.1 Ensuring informed consent
    - 10.2 Obtaining approval from the Institutional Review Board (IRB)
11. Time line
12. Budget and funding
13. Composition and roles of evaluation team

will be delivered at which stage of the analysis and to see whether the production of those results is on track with respect to the policy team's decision-making needs. Once the evaluation team has concluded the impact analysis, the initial results should be presented and shared with the policy team to ensure that any questions are answered and to prepare the dissemination stage.

Dissemination stage. In this stage, the policy team needs to ensure that the evaluation results reach the right people at the right time in an appropriate format. This is also the stage to ensure that all the data from the evaluation are appropriately documented. Often teams will engage multiple strategies and vehicles to disseminate results, keeping in mind the different target audiences, as discussed in chapter 14.

Establishing Collaboration

How to Set Up a Partnership

An evaluation is a balance between the technical expertise and independence contributed by the research team and the policy relevance, strategic guidance, and operational coordination contributed by the policy makers and program implementers on the policy team. A range of models can be used to set up and implement this partnership between the research and policy teams.

The choice of modality will depend on the context and objectives of the impact evaluation, as well as on the consideration of a range of risks. On the one hand, a fully independent research team with limited collaboration with the policy team may deliver an impact evaluation that is disconnected from the policy questions of interest or that uses a methodology constrained by insufficient interactions with program implementers. On the other hand, a research team fully integrated with the policy team may create risks of conflicts of interest or lead to the censorship of some results if open science principles are not applied (see chapter 13). In addition, evaluations can often have multiple goals, including building evaluation capacity within government agencies and sensitizing program operators to the realities of their projects as they are carried out in the field. These broader goals may also partly determine the model to be chosen.

Overall, what matters most for the quality of the impact evaluation is whether the partnership approach will produce unbiased estimates of program impacts. As long as principles of research ethics and open science are respected, unbiasedness and objectivity tend to be more critical to the quality of the impact evaluation than the functional independence of the research and policy teams. In practice, close collaboration between the research and policy teams is often needed to ensure that the highest-quality impact evaluation strategy is put in place.

The Outsourcing Model

For busy program implementers managing complex operations, vesting an outside team with the responsibility of designing and implementing the impact evaluation is often appealing. Outsourcing models can take different forms. Program managers sometimes outsource the design of the impact evaluation, as well as the implementation of the various surveys (typically, a baseline and follow-up survey), to a single entity in a wide-ranging contract. In other cases, program managers first outsource the design, and follow with contracts for various phases of data collection and analysis.

Outsourcing creates separation between the design and implementation of the impact evaluation, which can make the impact evaluation more independent. However, fully outsourcing the impact evaluation can come with substantial risks. The establishment of this kind of contractual relationship can limit the collaboration between the program implementation and contracted research teams.

In some cases, the contracted team is given a set of previously defined program parameters, with little margin to discuss design and implementation plans or the scope for shaping the research. In other cases, the program rules and implementation modalities needed to design a good impact evaluation may not be defined. In such cases, the contracted team in charge of the impact evaluation has limited influence in ensuring that these elements are defined.

In still other cases, the program may already have been designed or implementation may have begun, which can severely constrain methodological options for the evaluation. The contracted team is often asked to adjust to changes in program implementation ex post, without being closely involved or informed during implementation. These situations can lead to suboptimal evaluation designs or to challenges during implementation, as the contracted team may have different motivations from the researchers and policy makers who have led the design of the evaluation.

Lastly, the selection and oversight of the contracted team can be challenging for the program implementation unit. Procurement rules must be carefully considered up front to ensure that the outsourcing is efficient and does not present conflicts of interest. Certain rules may limit the possibility that a team contracted to contribute to the design of the impact evaluation can later bid on its implementation.

To mitigate these risks, it is generally preferable for the policy team to already have an impact evaluation design in place, including an identification strategy, core outcome indicators, initial power calculations, and approximate sample sizes. This will help guide the procurement and contracting, since these elements strongly affect evaluation budgets. The policy team should also establish mechanisms to ensure strong technical oversight of the design and implementation of the impact evaluation. This could be through an oversight committee or through regular technical and scientific review of impact evaluation products. Taken together, these mitigation measures suggest that the most effective model is usually not a full outsourcing model.

The Partnership Model

The collaboration between the research and policy teams is not necessarily built solely on contractual relationships. Mutually beneficial partnerships

can be put in place when researchers are interested in conducting research on a policy question and when policy makers and program implementers are seeking to ensure that a good-quality impact evaluation is set up in their project. Researchers have incentives to address new questions that will add to the global evidence base, and to push the envelope of the impact evaluation and contribute to its broader visibility. The research team may be able to leverage some cofinancing for the impact evaluation if the objectives of funders are closely aligned with the research focus of the evaluation.

Another type of integrated model that is becoming more prominent, especially in larger institutions, including the World Bank and the Inter-American Development Bank, uses in-house impact evaluation research capacity to support policy and program teams.

The partnership approach presents certain risks. At times, researchers may seek to incorporate novel research elements in the impact evaluation that may not be fully aligned to the immediate policy objectives at the local level, although they can add value more globally. For their part, policy makers and program implementers may not always appreciate the scientific rigor needed to undertake rigorous impact evaluations, and they may have a higher tolerance than the research team with respect to potential risks to the impact evaluation.

To mitigate those risks the objectives of the research team and policy teams need to be closely aligned. For instance, the research and policy teams can work together up front on a thorough evaluation plan outlining a detailed strategy as well as the respective teams' roles and responsibilities (see box 12.2). The impact evaluation plan is also a place to highlight key operational rules, as well as potential operational risks to the implementation of the impact evaluation.

A mutual commitment to an impact evaluation as embodied in a clear evaluation plan is essential for the partnership to work smoothly, even if a contractual relationship is not put in place. It is good practice for this mutual commitment to take the form of a written agreement—for instance, in the form of terms of reference or a memorandum of understanding—to establish the roles, responsibilities, and products of the impact evaluation. Such aspects can also be included in the impact evaluation plan.

The Fully Integrated Model

Some impact evaluations are implemented in a fully integrated model where the research and program implementation teams are one and the same. This approach is sometimes taken in efficacy trials, where new interventions are being tested for the *proof of concept*. In this case, researchers generally prefer to maintain control over implementation to ensure that the program is implemented as closely as possible to its original design.

While such impact evaluations are best able to test underlying theories and to establish whether a given intervention can work in ideal circumstances, the risk is that the results may have limited external validity.

Box 12.3 presents some examples of different models that research and policy teams can use to collaborate.

Box 12.3: Examples of Research–Policy Team Models

Outsourcing Evaluations at the Millennium Challenge Corporation

The Millennium Challenge Corporation (MCC), a U.S. aid agency, was established in 2004 with a strong emphasis on accountability and results. It requires each of its development programs to have a comprehensive monitoring and evaluation plan, with a focus on unbiased and independent evaluations. This focus led MCC to develop a model where both the design and implementation of evaluations are fully outsourced to external researchers. In the early years of MCC's operations, the separation between the program staff and the external researchers contracted for the evaluation sometimes created issues. For example, in Honduras, researchers designed a randomized controlled trial of a farmer training program. However, as the implementation contract was performance based, the implementer had a strong incentive to find high-performing farmers for the program. Eligible farmers were not randomly assigned into the program, invalidating the evaluation design. With the release of the first five evaluations of farmer training programs, MCC reflected on experiences like these and concluded that collaboration between implementers and evaluators is critical throughout design and implementation. The organization adapted its model to be more selective when applying impact evaluations in order to strike a balance between accountability and learning.

Integration at Innovations for Poverty Action

At Innovations for Poverty Action (IPA), a U.S.-based nonprofit organization, the researcher and policy teams work hand in hand from the very start of the evaluation design, and often from the time the program originates. IPA's model relies on an extensive network of field offices, many of which have existing relationships with government agencies and other implementing partners. From the time an evaluation is first conceived, IPA-affiliated researchers from a global network of universities work with country directors at relevant field offices to create an evaluation design and implementation plan. Country directors are responsible for leading partner relationships and matching principal investigators on the research team with program partners on the policy team to develop a proposal for an evaluation. Once a proposal has been approved, they hire project management staff to lead the data collection on the ground, all housed at the IPA field office. Coordination between the researchers and the program implementers is generally close, and in some cases, IPA offices are also responsible for implementing the intervention being evaluated.

Partnership Models at the World Bank

In the past decade, the World Bank has rapidly scaled up the use of prospective impact evaluations to assess the impacts

(continued)

Box 12.3: Examples of Research–Policy Team Models *(continued)*

of some of the development projects it finances. Several groups—including DIME (Development Impact Evaluation), SIEF (Strategic Impact Evaluation Fund), and GIL (Gender Innovation Lab)—provide funding and technical support to impact evaluations. When a particularly innovative or high-stakes project is put in place, impact evaluation activities are set up, either embedded in the project and managed by counterpart governments, or as independent activities managed by the World Bank. Either way, an evaluation team is put in place, consisting of a research team, including a mix of technical experts and academics, and a policy team, typically including policy makers, program implementers, and project operational team leaders.

For example, in Côte d'Ivoire, a joint initiative between the World Bank, the Abdul Latif Jameel Poverty Action Lab (J-PAL), and the government evaluated a Youth Employment and Skills Development Project. An evaluation team was put together, including a research team composed of a World Bank team leader, international academics, and local experts, and a policy team including specialists from the project implementing unit, the affiliated ministry, and World Bank staff. The evaluation team identified priority areas for impact evaluation. A prospective randomized controlled trial was put in place. The government shaped key questions and financed data collection, which was in part contracted out to the National School of Statistics (ENSEA) and partly conducted in-house by a dedicated data collection team. The World Bank financed technical oversight and research activities, as well as led the evaluation team. J-PAL contributed through affiliated academics. This model has proved effective in ensuring scientific rigor and global relevance, as well as alignment with policy makers' priorities. It requires careful management of partnerships and effective coordination across the various stakeholders in the evaluation team.

Sources: Bertrand and others 2016; IPA 2014; Sturdy, Aquino, and Molyneaux 2014.

Choosing a Research Team Partner

Policy makers and program implementers need to decide with whom to partner for the evaluation. Key questions are whether the research team—or parts of it—can be a local team, and what kind of outside assistance will be needed. Research capacity varies greatly from country to country. International firms are often contracted when particular skills are needed, and they can also partner with local firms. Data collection functions are generally implemented by local firms, given their deep knowledge of the local context and environment. There is also a strong global push to ensure the full participation of local researchers in impact evaluation.

As evaluation capacity increases, it is becoming more common for governments, private firms, and multilateral institutions to implement

impact evaluations in partnership with local research teams. Involving local researchers can bring significant value to the impact evaluation, given their knowledge of the local context. In some countries, research authorization is provided only to teams that include local researchers. Overall, it is up to the evaluation manager to assess local capacity and determine who will be responsible for what aspects of the evaluation effort. International impact evaluation networks of academics (such as J-PAL or IPA), private research firms, or impact evaluation groups in international institutions (such as DIME and SIEF at the World Bank, or SPD or RES at the Inter-American Development Bank) can help policy teams connect to international researchers with the technical expertise to collaborate on the impact evaluation.[1]

Another question is whether to work with a private firm or a public agency. Private firms or research institutions can be more dependable in providing timely results, but private firms often are understandably less amenable to incorporating elements into the evaluation that will make the effort costlier once a contract has been signed. The research team can also draw on research institutions and universities. Their reputation and technical expertise can ensure that evaluation results are widely accepted by stakeholders. However, those institutions sometimes lack the operational experience or the ability to perform some aspects of the evaluation, such as data collection. Such aspects may need to be subcontracted to another partner. Capacity building in the public sector may also be a goal and can be included as part of the terms of reference for the impact evaluation. Whatever combination of counterparts is finally crafted, a sound review of potential collaborators' past evaluation activities is essential to making an informed choice.

Particularly when working with a public agency with multiple responsibilities, the capacity and availability of an in-house research team to undertake the impact evaluation activities need to be assessed in light of other activities for which they are accountable. Awareness of the workload is important for assessing not only how it will affect the quality of the evaluation being conducted but also the opportunity cost of the evaluation with respect to other efforts for which the public agency is responsible.

How to Time the Evaluation

Part 1 discussed the advantages of prospective evaluations, designed during program preparation. Advance planning allows for a broader choice in generating comparison groups, facilitates the collection of baseline data, and

helps stakeholders reach consensus about program objectives and questions of interest.

Though it is important to plan evaluations early in the project design phase, carrying them out should be timed in the previously mentioned "sweet spot" once the program is mature enough to be stable, but before it is expanded. Pilot projects or nascent reforms are often prone to revision, both of their content and with respect to how, when, where, and by whom they will be implemented. Program providers may need time to learn and consistently apply new operational rules. Because evaluations require clear rules of program operation to generate appropriate counterfactuals, it is important to apply evaluations to programs after they are well established.

Another key issue concerns how much time is needed before results can be measured. The right balance is context-specific: "If one evaluates too early, there is a risk of finding only partial or no impact; too late, and there is a risk that the program might lose donor and public support or that a badly designed program might be expanded" (King and Behrman 2009, 56).[2] A range of factors needs to be weighted to determine when to collect follow-up data:

The program cycle, including program duration, time of implementation, and potential delays. The impact evaluation needs to be fitted to the program implementation cycle; the evaluation cannot drive the program being evaluated. By their very nature, evaluations are subject to the program time frame; they must be aligned to the expected duration of the program. They also must be adapted to potential implementation lags when programs are slow to assign benefits or are delayed by external factors.[3] In general, although evaluation timing should be built into the project from the outset, evaluators should be prepared to be flexible and to make modifications as the project is implemented. In addition, provision should be made for tracking the interventions, using a strong monitoring system so that the evaluation effort is informed by the actual pace of the intervention.

The expected time needed for the program to affect outcomes, as well as the nature of outcomes of interest. The timing of follow-up data collection must take into account how much time is needed after the program is implemented for results to become apparent. The program results chain helps identify outcome indicators and the appropriate time to measure them. Some programs (such as income support programs) aim to provide short-term benefits, whereas others (such as basic education programs) aim for longer-term gains. Moreover, certain results by their nature take longer to appear (such as changes in life expectancy or fertility from a health reform) than others (such as earnings from a training program).

For example, in the evaluation of the Bolivian Social Investment Fund, which relied on baseline data collected in 1993, follow-up data were not collected until 1998 because of the time required to carry out the interventions (water and sanitation projects, health clinics, and schools) and for effects on the beneficiary population's health and education to emerge (Newman and others 2002). A similar period of time was required for the evaluation of a primary education project in Pakistan that used an experimental design with baseline and follow-up surveys to assess the impact of community schools on student outcomes, including academic achievement (King, Orazem, and Paterno 2008). However, follow-up data are often collected earlier than would be recommended, given pressures for timely results or budget and project cycle constraints (McEwan 2014).

When to collect follow-up data will therefore depend on the program under study, as well as on the outcome indicators of interest.

Follow-up data can be collected more than once, so that short-term and medium-term results can be considered and contrasted while the treatment group is still receiving the intervention. Follow-up data may not capture the full impact of the program if indicators are measured too early. Still, it is very useful to document short-term impacts, which can provide information about expected longer-term outcomes to produce early impact evaluation results that can invigorate dialogue between the research and policy teams and maintain contact with the evaluation sample to reduce sample attrition over time.

Follow-up surveys that measure long-term outcomes after the program has been implemented often produce the most convincing evidence regarding program effectiveness. For instance, the positive results from long-term impact evaluations of early childhood programs in the United States (Currie 2001; Currie and Thomas 1995, 2000) and Jamaica (Grantham-McGregor and others 1994; Gertler and others 2014) have been influential in making the case for investing in early childhood interventions.

Long-term impacts sometimes are explicit program objectives, but even a strong impact evaluation design may not withstand the test of time. For example, units in the control group may begin to benefit from spillover effects from program beneficiaries.

Teams can collect follow-up data more than once, so that short-, medium-, and long-term results can be considered and contrasted.

Policy-making cycles. The timing of an evaluation must also take into account when certain information is needed to inform policy decisions and must synchronize evaluation and data collection activities to key decision-making points. The production of results should be timed to inform budgets, program expansion, or other policy decisions.

How to Budget for an Evaluation

Budgeting constitutes one of the last steps to operationalize the evaluation design. In this section, we review some existing impact evaluation cost data, discuss how to budget for an evaluation, and suggest some options for funding.

Review of Cost Data

Tables 12.1 and 12.2 provide useful benchmarks on the costs associated with conducting rigorous impact evaluations. They contain cost data on impact evaluations of a number of projects supported by the Strategic Impact Evaluation Fund (SIEF) administered by the World Bank. The sample in table 12.1 comes from a comprehensive review of programs supported by the Early Childhood Development and Education research clusters within SIEF. The sample in table 12.2 was selected based on the availability of current budget statistics from the set of impact evaluations financed by SIEF.[4]

Key Concept

Impact evaluations generally constitute only a small percentage of overall program budgets. In addition, the cost of conducting an impact evaluation must be compared with the opportunity costs of not conducting a rigorous evaluation and thus potentially running an ineffective program.

The direct costs of the evaluation activities reviewed in the samples included in tables 12.1 and 12.2 range between US$130,000 and US$2.78 million, with an average cost of about US$1 million. Although those costs vary widely and may seem high in absolute terms, impact evaluations generally constitute only a small percentage of overall program budgets. In addition, the cost of conducting an impact evaluation must be compared with the opportunity costs of not conducting a rigorous evaluation and thus potentially running an ineffective program. Evaluations allow researchers and policy makers to identify which programs or program features work, which do not, and which strategies may be the most effective and efficient in achieving program goals. In this sense, the resources needed to implement an impact evaluation constitute a relatively small but significant investment.

Table 12.2 disaggregates the costs of the sample of impact evaluations supported by SIEF. The total costs of an evaluation include World Bank staff time, national and international consultants, travel, data collection, and dissemination activities.[5] As is the case in almost all evaluations for which existing data cannot be used, the highest cost in the evaluation is new data collection, accounting for 63 percent of the total evaluation cost, on average, as shown in the table.

These numbers reflect different sizes and types of evaluations. The relative cost of evaluating a pilot program is generally higher than the relative cost of evaluating a nationwide or universal program. In addition, some evaluations require only one follow-up survey or may be able to use existing data sources, whereas others may need to carry out multiple rounds of data collection. Costs of data collection depend largely on the salaries of the local

Table 12.1 Cost of Impact Evaluations of a Selection of World Bank–Supported Projects

Impact evaluation (IE)	Country	Total cost of impact evaluation (US$)	Total cost of program[a] (US$)	IE costs as a percentage of total program costs
Safety net project	Burkina Faso	750,000	38,800,000	1.9
Migrant Skills Development and Employment	China	220,000	50,000,000	0.4
Social Safety Net Project	Colombia	130,000	86,400,000	0.2
Integrated Nutrition/Workfare Social Safety Net (Pilot)	Djibouti	480,000	5,000,000	8.8
Social Sectors Investment Program	Dominican Republic	600,000	19,400,000	3.1
Performance-Based Incentives for Teachers	Guinea	2,055,000	39,670,000	4.9
Social Protection	Jamaica	800,000	40,000,000	2.0
Addressing Chronic Malnutrition	Madagascar	651,000	10,000,000	6.1
Community-Based Childcare Centers (pilot)	Malawi	955,000	1,500,000	38.9
Information and Unconditional Cash Transfer	Nepal	984,000	40,000,000	2.4
Social Safety Net Technical Assistance	Pakistan	2,000,000	60,000,000	3.3
Social Protection Project	Panama	1,000,000	24,000,000	4.2
1st Community Living Standards	Rwanda	1,000,000	11,000,000	9.1
Information-for-accountability and teacher incentive interventions	Tanzania	712,000	416,000,000	0.2
Class-size and teacher quality interventions	Uganda	639,000	100,000,000	0.6
Social Fund for Development 3	Yemen, Rep.	2,000,000	15,000,000	13.3
Average		**936,000**	**59,798,000**	**6.2**

Source: A sample of impact evaluations supported by the Early Childhood Development and Education research clusters of the World Bank's Strategic Impact Evaluation Fund.

Note: IE = impact evaluation.

a. Total cost of program does not include costs associated with the impact evaluation.

team, the cost of accessing populations in the evaluation sample, and the length of time in the field. To learn more about how to estimate the cost of a survey in a particular context, it is recommended that the evaluation team first contact the national statistical agency and look for information from teams who have done survey work in the country.

Table 12.2 Disaggregated Costs of a Selection of World Bank–Supported Impact Evaluations

Impact evaluation	Country	Total cost[a] (US$)	Sample size	Data collection (percent)[b]	Staff and consultants (percent)[b]	Travel (percent)[b]	Dissemination and workshops (percent)[b]	Other (percent)[b]
Building Parental Capacity to Help Child Nutrition and Health	Bangladesh	655,000	2,574 households	27	48	5	0	20
Closing the Early Learning Gap for Roma Children	Bulgaria	702,000	6,000 households	74	21	4	1	0
The ECD and Nutrition Component of Burkina Faso's Safety Net Project	Burkina Faso	750,000	4,725 households	55	20	3	1	21
Payment of Community Teachers	Chad	1,680,000	2,978 schools	52	14	12	18	4
A Home-Based Early Childhood Development Intervention	Colombia	573,000	1,429 individuals	54	36	2	2	7
Piloting an Integrated Nutrition/Workfare Social Safety Net	Djibouti	480,000	1,150 individuals	75	0	0	6	18
Supervision and Incentives for Increased Learning: The TCAI High Performance Program	Ghana	498,000	480 schools	51	46	3	0	0
Performance-Based Incentives for Teachers	Guinea	2,055,000	420 schools	82	9	3	1	4

(continued)

Table 12.2 *(continued)*

Impact evaluation	Country	Total cost[a] (US$)	Sample size	Data collection (percent)[b]	Staff and consultants (percent)[b]	Travel (percent)[b]	Dissemination and workshops (percent)[b]	Other (percent)[b]
Education Service Delivery Support	Haiti	436,000	200 schools	40	31	17	3	9
Non-financial Extrinsic and Intrinsic Teacher Motivation	India	448,000	360 schools	83	5	11	1	0
Early Childhood Stimulation and Social Accountability in India's Integrated Child Development Strategy	India	696,000	2,250 individuals	49	43	5	3	0
Women's Self-help Groups to Strengthen Health, Nutrition, Sanitation, and Food Security	India	844,000	3,000 households	52	39	5	1	2
Early Childhood Development for the Poor	India	1,718,000	2,588 households	46	53	1	1	0
Early Childhood Nutrition, Availability of Health Service Providers, and Life Outcomes as Young Adults	Indonesia	2,490,000	6,743 individuals	94	0	2	4	0
Addressing Chronic Malnutrition	Madagascar	651,000	5,000 individuals	0	0	66	2	32
Integrated Parenting, Nutrition, and Malaria Prevention	Mali	949,000	3,600 individuals	58	22	4	5	11

(continued)

Table 12.2 *(continued)*

Impact evaluation	Country	Total cost[a] (US$)	Sample size	Data collection (percent)[b]	Staff and consultants (percent)[b]	Travel (percent)[b]	Dissemination and workshops (percent)[b]	Other (percent)[b]
Increasing Education Accountability through Community-Based Pedagogical Assistants	Mexico	268,000	230 schools	70	26	3	2	0
Access to a Private Comprehensive Schooling Model	Mexico	420,000	172 individuals	45	48	5	1	1
Randomized Impact Evaluation of Various Early Literacy Reading Skills Interventions	Mozambique	1,762,000	110 schools	78	5	4	8	6
Integrated Early Childhood Development and Nutrition	Mozambique	1,908,000	6,700 households	74	8	5	7	7
A Health Insurance Pilot Program	Nepal	485,000	6,300 households	61	33	3	4	0
Information and Unconditional Cash Transfers on Nutritional Outcomes	Nepal	984,000	3,000 individuals	57	23	9	1	10
Cash Transfers, Parenting Training, and Holistic Early Childhood Development	Niger	984,000	4,332 households	67	18	7	1	7
Understanding the Dynamics of Information for Accountability	Nigeria	1,052,000	120 schools	59	25	8	3	6

(continued)

Table 12.2 *(continued)*

Impact evaluation	Country	Total cost[a] (US$)	Sample size	Data collection (percent)[b]	Staff and consultants (percent)[b]	Travel (percent)[b]	Dissemination and workshops (percent)[b]	Other (percent)[b]
Subsidy Reinvestment and Empowerment Programme and Maternal and Child Health Initiative	Nigeria	2,775,000	5,000 households	76	13	6	4	2
Community Engagement for School Committee	Pakistan	845,000	287 schools	59	15	6	3	18
Strengthening Private Schools for the Rural Poor	Pakistan	2,124,000	2,000 schools	26	25	5	2	42
Selection and Motivational Impacts of Performance Contracts for Primary School Teachers	Rwanda	797,000	300 schools	79	7	3	1	11
Information Campaign in Primary Schools	South Africa	647,000	200 schools	67	24	2	3	4
Testing Information for Accountability and Teacher Incentive Interventions	Tanzania	712,000	420 schools	86	6	7	2	0
Designing Effective Teacher Incentive Programs	Tanzania	889,000	420 schools	85	11	2	2	0
Program for Women at High Risk of HIV Infection	Tanzania	1,242,000	3,600 individuals	90	7	2	1	0

(continued)

Table 12.2 *(continued)*

Impact evaluation	Country	Total cost[a] (US$)	Sample size	Data collection (percent)[b]	Staff and consultants (percent)[b]	Travel (percent)[b]	Dissemination and workshops (percent)[b]	Other (percent)[b]
Class-Size and Teacher Quality Interventions	Uganda	639,000	200 schools	82	9	7	2	0
Contrasting Efficiency of Education Service Delivery in Public and Private Sectors	Uganda	737,000	280 schools	77	18	3	3	0
Average		**1,026,000**		**63**	**21**	**7**	**3**	**7**

Source: A sample of impact evaluations financed by the World Bank's Strategic Impact Evaluation Fund.

a. Estimated costs do not always capture the full costs of the evaluation, including the time of the policy team.

b. Percent of total cost of the evaluation by category. This cost does not include the costs of local project staff, who were often heavily engaged in the design and supervision of the evaluation, as accurate data on these costs are not regularly recorded.

Budgeting for an Impact Evaluation

Many resources are required to implement a rigorous impact evaluation, especially when primary data are being collected. Budget items include staff fees for at least one principal investigator/researcher, a field coordinator, a sampling expert, and a data collection team. Time from project staff on the policy team is also needed to provide guidance and support throughout the evaluation. These human resources may consist of researchers and technical experts from international organizations, international or local consultants, and local program staff. The costs of travel and subsistence must also be budgeted. Resources for dissemination, often in the form of workshops, reports, and academic papers, should be considered in the evaluation planning.

As noted, the largest costs in an evaluation are usually those of data collection (including creating and pilot testing the survey), data collection materials and equipment, training for the enumerators, daily wages for the enumerators, vehicles and fuel, and data entry operations. Calculating the costs of all these inputs requires making some assumptions about, for example, how long the questionnaire will take to complete and travel times between sites.

The costs of an impact evaluation may be spread out over several years. A sample budget in table 12.3 shows how the expenditures at each stage of an evaluation can be disaggregated by year for accounting and reporting purposes. Again, budget demands will likely be higher during the years when the data are collected.

Options for Funding Evaluations

Financing for an evaluation can come from many sources, including project resources, direct program budgets, research grants, or donor funding. Often, evaluation teams look to a combination of sources to generate the needed funds. Although funding for evaluations used to come primarily from research budgets, a growing emphasis on evidence-based policy making has increased funding from other sources. In cases where an evaluation is likely to fill a substantial knowledge gap that is of interest to the development community more broadly, and where a credible, robust evaluation can be applied, policy makers should be encouraged to look for outside funding, given the public good the evaluation results will provide. Sources of funding include the government, development banks, multilateral organizations, United Nations agencies, foundations, philanthropists, and research and evaluation organizations such as the International Initiative for Impact Evaluation.

Table 12.3 Sample Budget for an Impact Evaluation

	Design stage				Baseline data stage			
	Unit	Cost per unit (US$)	No. of units	Total cost (US$)	Unit	Cost per unit (US$)	No. of units	Total cost (US$)
A. Staff salaries	Weeks	7,500	2	15,000	Weeks	7,500	2	15,000
B. Consultant fees				14,250				41,900
International consultant (1)	Days	450	15	6,750	Days	450	0	0
International consultant (2)	Days	350	10	3,500	Days	350	10	3,500
Research assistant/field coordinator	Days	280	0	0	Days	280	130	36,400
Statistical expert	Days	400	10	4,000	Days	400	5	2,000
C. Travel & subsistence								
Staff: International airfare	Trips	3,350	1	3,350	Trips	3,350	1	3,350
Staff: Hotel & per diem	Days	150	5	750	Days	150	5	750
Staff: Local ground transport	Days	10	5	50	Days	10	5	50
International consultants: International airfare	Trips	3,500	2	7,000	Trips	3,500	2	7,000
International consultants: Hotel & per diem	Days	150	20	3,000	Days	150	20	3,000
International consultants: Local ground transport	Days	10	5	50	Days	10	5	50
Field coordinator: International airfare	Trips		0	0	Trips	1,350	1	1,350

(continued)

Table 12.3 *(continued)*

	Design stage				Baseline data stage			
	Unit	**Cost per unit (US$)**	**No. of units**	**Total cost (US$)**	**Unit**	**Cost per unit (US$)**	**No. of units**	**Total cost (US$)**
Field coordinator: Hotel & per diem	Days		0	0	Days	150	3	150
Field coordinator: Local ground transport	Days		0	0	Days	10	3	30
D. Data collection								126,000
Data type 1: Consent					School	120	100	12,000
Data type 2: Education outcomes					Child	14	3,000	42,000
Data type 3: Health outcomes					Child	24	3,000	72,000
E. Data analysis and dissemination								
Workshop(s)								
Dissemination/reporting								
Total costs per stage	Design stage			43,450	Baseline stage			198,630

(continued)

Table 12.3 *(continued)*

	Follow-up data, Stage I				Follow-up data, Stage II			
	Unit	**Cost per unit (US$)**	**No. of units**	**Total cost (US$)**	**Unit**	**Cost per unit (US$)**	**No. of units**	**Total cost (US$)**
A. Staff salaries	Weeks	7,500	22222	15,000	Weeks	7,500	22	15,000
B. Consultant fees				43,750				38,000
International consultant (1)	Days	450	15	6,750	Days	450	10	4,500
International consultant (2)	Days	350	20	7,000	Days	350	10	3,500
Research assistant/field coordinator	Days	280	100	28,000	Days	280	100	28,000
Statistical expert	Days	400	5	2,000	Days	400	5	2,000
C. Travel & subsistence								
Staff: International airfare	Trips	3,350	1	3,350	Trips	3,350	2	6,700
Staff: Hotel & per diem	Days	150	10	1,500	Days	150	10	1,500
Staff: Local ground transport	Days	10	5	50	Days	10	5	50
International consultants: International airfare	Trips	3,500	2	7000	Trips	3,500	2	7000
International consultants: Hotel & per diem	Days	150	20	3,000	Days	150	20	3,000
International consultants: Local ground transport	Days	10	5	50	Days	10	5	50
Field coordinator: International airfare	Trips	1,350	1	1,350	Trips	1,350	1	1,350

(continued)

Table 12.3 *(continued)*

	Follow-up data, Stage I				Follow-up data, Stage II			
	Unit	**Cost per unit (US$)**	**No. of units**	**Total cost (US$)**	**Unit**	**Cost per unit (US$)**	**No. of units**	**Total cost (US$)**
Field coordinator: Hotel & per diem	Days	150	3	450	Days	150	3	450
Field coordinator: Local ground transport	Days	10	3	30	Days	10	3	30
D. Data Collection				126,000				126,000
Data type 1: Consent	School	120	100	12,000	School	120	100	12,000
Data type 2: Education outcomes	Child	14	3,000	42,000	Child	14	3,000	42,000
Data type 3: Health outcomes	Child	24	3,000	72,000	Child	24	3,000	72,000
E. Data analysis and dissemination								55,000
Workshop(s)						20,000	2	40,000
Dissemination/reporting						5,000	3	15,000
Total costs per stage	Follow-up, Stage I			201,530	Follow-up, Stage II			254,130
					Total evaluation costs			697,740

Additional Resources

- For accompanying material to this chapter and hyperlinks to additional resources, please see the Impact Evaluation in Practice website (http://www.worldbank.org/ieinpractice).
- To access several tools to help plan and implement an evaluation, see the Inter-American Development Bank Evaluation portal (http://www.iadb.org/evaluationhub), including the following:
  - Design section: Gantt charts to assist in the scheduling of impact evaluation activities, a budget template tool to estimate the costs of an impact evaluation, and a checklist of core activities to be carried out.
  - Implementation section: Sample terms of reference (TORs) for principal investigators, data collection firms, and technical support and supervision.
- For guidelines and tools to help plan and implement an evaluation, see the World Bank Impact Evaluation Toolkit (Vermeersch, Rothenbühler, and Sturdy 2012), including the following:
  - Module 2: Team Building: Sample terms of reference for principal investigators, evaluation coordinators, data analysts, local researchers, power calculation experts, data quality experts, field workers, and others.
  - Field manuals and training programs for household and health facilities.
  - Module 3: Design: Guidelines on how to align the timing, team composition, and budget of your impact evaluation; and a budget template.
  - Module 4: Data Collection Preparation: Information on scheduling data collection activities and reaching agreements with stakeholders on data ownership; Gantt chart; sample data collection budget.

Notes

1. J-PAL is the Abdul Latif Jameel Poverty Action Lab. SPD is the Inter-American Development Bank's (IDB) Office of Strategic Planning and Development Effectiveness. RES is IDB's Research Department.
2. For a detailed discussion of timing issues in relation to the evaluation of social programs, see King and Behrman (2009).
3. "There are several reasons why implementation is neither immediate nor perfect, why the duration of exposure to a treatment differs not only across program areas but also across ultimate beneficiaries, and why varying lengths of exposure might lead to different estimates of program impact" (King and Behrman 2009, 56).
4. While tables 12.1 and 12.2 provide useful benchmarks, they are not representative of all evaluations undertaken by the SIEF program or the World Bank.
5. In this case, cost is calculated as a percentage of the portion of the project cost financed by the World Bank.

References

Bertrand, Marianne, Bruno Crépon, Alicia Marguerie, and Patrick Premand. 2016. "Impacts à Court et Moyen Terme sur les Jeunes des Travaux à Haute Intensité de Main d'oeuvre (THIMO) : Résultats de l'évaluation d'impact de la composante THIMO du Projet Emploi Jeunes et Développement des compétence (PEJEDEC) en Côte d'Ivoire." Washington, DC: Banque Mondiale et Abidjan, BCP-Emploi.

Currie, Janet. 2001. "Early Childhood Education Programs." *Journal of Economic Perspectives* 15 (2): 213–38.

Currie, Janet, and Duncan Thomas. 1995. "Does Head Start Make a Difference?" *American Economic Review* 85 (3): 341–64.

———. 2000. "School Quality and the Longer-Term Effects of Head Start." *Journal of Economic Resources* 35 (4): 755–74.

Gertler, Paul, James Heckman, Rodrigo Pinto, Arianna Zanolini, Christel Vermeersch, and others. 2014. "Labor Market Returns to an Early Childhood Stimulation Intervention in Jamaica." *Science* 344 (6187): 998–1001.

Grantham-McGregor, Sally, Christine Powell, Susan Walker, and John Himes. 1994. "The Long-Term Follow-up of Severely Malnourished Children Who Participated in an Intervention Program." *Child Development* 65: 428–93.

IPA (Innovations for Poverty Action). 2014. "Researcher Guidelines: Working with IPA." September 1. http://www.poverty-action.org/sites/default/files/researcher_guidelines_version_2.0.pdf.

King, Elizabeth M., and Jere R. Behrman. 2009. "Timing and Duration of Exposure in Evaluations of Social Programs." *World Bank Research Observer* 24 (1): 55–82.

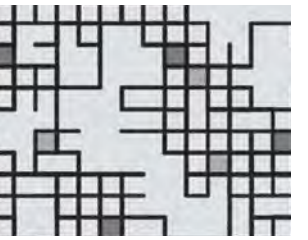
King, Elizabeth M., Peter F. Orazem, and Elizabeth M. Paterno. 2008. "Promotion with and without Learning: Effects on Student Enrollment and Dropout Behavior." Policy Research Working Paper 4722, World Bank, Washington, DC.

McEwan, Patrick J. 2014. "Improving Learning in Primary Schools of Developing Countries: A Meta-Analysis of Randomized Experiments." *Review of Educational Research*. doi:10.3102/0034654314553127.

Newman, John, Menno Pradhan, Laura B. Rawlings, Geert Ridder, Ramiro Coa, and Jose Luis Evia. 2002. "An Impact Evaluation of Education, Health, and Water Supply Investments by the Bolivian Social Investment Fund." *World Bank Economic Review* 16 (2): 241–74.

Sturdy, Jennifer, Sixto Aquino, and Jack Molyneaux. 2014. "Learning from Evaluation at the Millennium Challenge Corporation." *Journal of Development Effectiveness* 6 (4): 436–50.

Vermeersch, Christel, Elisa Rothenbühler, and Jennifer Sturdy. 2012. *Impact Evaluation Toolkit: Measuring the Impact of Results-based Financing on Maternal and Child Health*. World Bank, Washington, DC. http://www.worldbank.org/health/impactevaluationtoolkit.



CHAPTER 13

The Ethics and Science of Impact Evaluation

Managing Ethical and Credible Evaluations

The ethics of evaluation center on protecting the individuals, or human subjects, who participate in the evaluation, while transparency of methods helps ensure that the results of the evaluation are unbiased, reliable, and credible, and contribute to a wider body of knowledge.

Policy makers and researchers have a joint interest and responsibility to ensure that the evaluation is ethical and that its results are unbiased, reliable, and credible. Failure to do so can invalidate the evaluation and lead to problems beyond the scope of the evaluation. Imagine an impact evaluation that endangers a group of people by releasing personal data, or an evaluation that uses a program assignment mechanism that is unfair by excluding the neediest families. Imagine an evaluation that shows that a program is highly successful, but doesn't make any data available to support the claim. Any of these cases could lead to public outcry; to complaints in the media, in courts, or elsewhere; and to embarrassment for policy makers and researchers alike. Criticism of the evaluation might spill over to the program itself and even undermine its implementation. Reliability and completeness of evaluation results are also very important: when evaluations produce biased or partial

estimates of the impact of programs, policy makers will be restricted in their ability to make a fully informed decision.

While impact evaluations are linked to public programs and projects, they are also a research activity and thus are conducted in the realm of social science. Accordingly, the evaluation team must abide by a number of social science principles and rules to make sure the evaluation is ethical and transparent in its methods and results.

The Ethics of Running Impact Evaluations

When an impact evaluation assigns subjects to treatment and comparison groups and collects and analyzes data about them, the evaluation team has a responsibility to minimize to the greatest extent possible any risks that individuals might be harmed and to ensure that those individuals who participate in the evaluation are doing so through informed consent.

The Ethics of Assignment to Treatment and Comparison Groups

As with the Hippocratic Oath in the medical profession, a first principle of evaluation ethics should be to do no harm. The foremost concern is that the program intervention to be evaluated might harm individuals, either directly or indirectly. For example, a road rehabilitation project might displace households living along some sections of the roads. Or a literacy project that does not take into account the use of native languages might harm indigenous communities. Many governments and international donors that finance development projects use a safeguards framework to prevent and mitigate these types of risks. While the program implementers have the primary responsibility to apply project safeguard measures, the evaluation team should be vigilant to verify that the project is complying with these required frameworks.

There is another concern about harm that may arise from withholding an intervention from potential beneficiaries. A fundamental principle is that groups should not be excluded from an intervention that is known to be beneficial solely for the purpose of conducting an evaluation. Evaluations should only be done in cases where the evaluation team does not know whether an intervention is beneficial in the particular context where it is being evaluated. Additionally, if an evaluation shows that a program is cost-effective, the funders of the program—whether governments, donors, or nongovernmental organizations—should make reasonable efforts to expand the program to include the comparison groups once the impact evaluation has been completed.

A related principle that we advocate in this book is that evaluations should not dictate how programs are assigned; instead, evaluations should be fitted to program assignment rules to the extent that those are clear and fair. The evaluation can also help (re)define rules when they don't exist or when they are not fair. Following this procedure will help ensure that ethical concerns will not stem so much from the impact evaluation itself but rather from the ethics of the rules used to choose the beneficiaries of the program. Nonetheless, the assignment into treatment and comparison groups can raise concerns about the ethics of denying program benefits to eligible beneficiaries. This is particularly the case with randomized assignment of program benefits. In part 2 and in chapter 11, we have emphasized that randomized assignment is a method that can be applied in specific operational contexts. In particular, the fact that most programs operate with limited financial and administrative resources makes it impossible to reach all eligible beneficiaries at once. This addresses the ethical concerns, since the program itself must develop allocation rules and impose some form of rationing, even without the existence of an impact evaluation. From an ethical standpoint, there is a good case to be made for all of those who are equally eligible to participate in a program to have the same chance of receiving the program. Randomized assignment fulfills this requirement. In other operational contexts where a program will be phased in over time, rollout can be based on randomly selecting the order in which equally deserving beneficiaries or groups of beneficiaries will receive the program. Again, this will give each eligible beneficiary the same chance to be the first to receive the program. In these cases, beneficiaries who enter the program later can be used as a comparison group for earlier beneficiaries, generating a solid evaluation design, as well as a transparent and fair method for allocating scarce resources.

Key Concept

Groups should not be excluded from an intervention that is known to be beneficial solely for the purpose of an evaluation.

Finally, there can also be an ethical concern about *not* pursuing an evaluation when programs invest substantial resources in interventions whose effectiveness is unknown. In this context, the lack of evaluation could itself be seen as unethical because it might perpetuate wasteful programs that do not benefit the population, while the funds might be better spent on more effective interventions. The information about program effectiveness that impact evaluations yield can lead to more effective and ethical investment of public resources.

Protecting Human Subjects during Data Collection, Processing, and Storage

A second point at which subjects might be harmed is during data collection, processing, and storage. The households, teachers, doctors, administrators, and others who respond to questionnaires or provide data through other

means are subject to harm if the information they provide is made publicly available without sufficient safeguards to protect their anonymity. The harm might affect the individuals themselves or an organization to which they belong. Here are a few examples:

- While a survey is being administered, a woman shares information about her family planning practices, and her husband (who does not favor family planning) overhears her conversation with the enumerator.
- The privacy of households is violated (and their safety is jeopardized) when an individual manages to use survey data that were posted on the Internet to identify the income and assets of specific families.
- A study uses inappropriately qualified enumerators to administer biometric tests, such as drawing blood.
- A survey respondent asks to withdraw from a study halfway through the interview but is instructed by the enumerator to finish answering the survey questions.
- Survey data are used to identify community organizations that oppose certain government policies, and to retaliate against them.

Given risks like these, it is the responsibility of the principal investigators and others on the research team to safeguard the rights and welfare of human subjects involved in the impact evaluation in accordance with the appropriate national code of ethics or legislation and with international guidines.[1] The World Health Organization (WHO) recommends the following basic criteria for assessing the research projects involving human subjects:

- The rights and welfare of the subjects involved in the impact evaluation should be adequately protected.
- The researchers should obtain freely given, informed consent from the participants.
- The balance between risk and potential benefits involved should be assessed and deemed acceptable by a panel of independent experts.
- Any special national requirements should be met.

The *Belmont Report: Ethical Principles and Guidelines for the Protection of Human Subjects of Research* (National Commission 1978) identifies three principles that form the foundation for the ethical conduct of research involving human subjects:

- *Respect for persons*. How will the researchers obtain informed consent from their research subjects?

- *Beneficence*. How will the researchers ensure that the research (1) does not harm and (2) maximizes potential benefits and minimizes potential harm?
- *Justice*. How will the researchers ensure that the benefits and burdens of research are fairly and equitably shared?

As a key element of his or her duty to protect human subjects, the principal investigator(s) should submit the research and data collection protocols for review and clearance to an institutional review board (IRB), also known as an independent ethics committee (IEC) or ethical review board (ERB). The IRB is a committee that has been formally designated to review, approve, and monitor biomedical and behavioral research involving human subjects. Both before the study starts and during its implementation, the IRB reviews the research protocols and related materials in order to assess the ethics of the research and its methods. In the context of impact evaluations, IRB review is particularly important when the study requires the collection of household and individual data. In particular, the IRB review checks whether the participants are capable of making the choice to participate in the data collection activities and whether their choice will be fully informed and voluntary. Finally, the IRB reviews whether there is any reason to believe that the safety of participants could be at risk.

The principal investigator is responsible for identifying all the institutions that should review and clear the study. Many countries have a national ethical review board, and most universities have an institutional review board. Typically, the team will be required to obtain ethical clearance from the respective country's national ethical review board and from the institutional review board of any university with which the investigators are affiliated. There may be particular instances where impact evaluations are carried out in countries that do not have a national ethical review board, or with researchers whose institutions do not have an institutional review board. In those cases, the principal investigator should contract a third-party (possibly commercial) institutional review board. The review and clearance process can take two to three months, though the time varies depending on how often the IRB committee meets. The policy and research team should coordinate submissions to the IRB and data collection activities so that they can obtain all required clearances before initiating data collection that involves human subjects.

Review by an IRB is a necessary but insufficient condition to ensure human subjects' protection. IRBs can vary widely in their capacity and experience with social science experiments, as well as in the focus of their review. IRBs, especially if their location is far away from where the evaluation is taking place, may be insufficiently aware of local circumstances to be able to identify contextual threats to human subjects. They may put

excessive emphasis on the wording of questionnaires and consent forms. Or they may have experience in a more focused subject area, such as medical experiments, whose norms are quite different from social experiments in terms of risks to human subjects. Thinking about human subject protection doesn't stop once IRB approval is obtained; rather, it should be seen as a starting point for ensuring that the evaluation is ethical.

Key Concept

An institutional review board (IRB) is a committee that has been designated to review, approve, and monitor research involving human subjects.

Institutional review boards commonly require the following information to be submitted for review:

Evidence of training. Many IRBs (as well as many national ethical guidelines) require that the research team be trained in the protection of human subjects, though modalities vary by country. We list several options for training in the additional resources section at the end of this chapter.

The research protocol. The research protocol includes core elements usually outlined in the evaluation plan—notably the purpose of the study and objectives of the evaluation, core policy questions, and proposed evaluation methodology—as well as description of how the research team will ensure that human subjects are protected. As such, it is an important document in an evaluation's documentation. The research protocol normally includes the following elements with respect to the treatment of human subjects: the criteria for selecting study participants (subjects), the methodology and protocols applied for protecting vulnerable subjects, procedures used to ensure that subjects are aware of the risks and benefits of participation in the study, and procedures used to ensure anonymity. The research protocol should be used by the survey firm to guide fieldwork procedures. More information on the content of the research protocol can be found on the World Health Organization (WHO) website and in the Impact Evaluation Toolkit.[2]

Procedures for requesting and documenting informed consent. Informed consent is one of the cornerstones of protecting the rights of human subjects in any study. It requires that respondents have a clear understanding of the purpose, procedures, risks, and benefits of the data collection in which they are asked to participate. By default, informed consent by an adult respondent requires a written document that includes a section on the methods used to protect respondent confidentiality, a section on the respondent's right to refuse or cease participation at any point in time, an explanation of potential risks and benefits, contact information in the event the respondent wishes to contact the data collection team, and space for respondents to record their formal written consent to participate in the data collection with a signature. Sometimes, study participants are not capable of making the choice to participate. For example, children are usually deemed to be incapable of making this choice. Therefore, in contrast to able adults, minors cannot consent to participate in a survey; they may assent to participate after written permission by their parent or guardian. While the steps described

are the default informed procedures, many impact evaluations request one or more waivers from their IRB from the requirement to obtain formal written consent from respondents. For example, when dealing with an illiterate population, formal written consent among eligible, potential adult respondents is often waived and replaced with documented verbal consent.[3]

Procedures for protecting respondent confidentiality. Protection of respondent confidentiality is critical when storing data and making data publically available. All information provided during the course of data collection should be anonymized to protect the identity of the respondents. Although results of the study may be published, the report should be written in such a way that it is not possible to identify an individual or household. With respect to ensuring confidentiality in the data, each subject of the survey should be assigned a unique encrypted identification number (ID), and all names and identifiers should be deleted from the database that is made publicly available. Identifiers include any variables allowing identification of individuals or households (such as address), or any combination of variables that does the same (such as a combination of date of birth, place of birth, gender, and years of education). In case the research team anticipates that it would need the identifiers in order to follow up on respondents in a subsequent survey, it can keep a separate and securely kept database that links the encrypted IDs with the respondents' identifying information.[4] In addition to encrypting individual IDs, it may also be necessary to encrypt locations and institutions. For example, if households and individuals are coded with encrypted IDs but villages are identified, it might still be possible to identify households through the characteristics that are included in the survey. For example, a particular village may include only one household that owns a motorcycle, seven cows, and a barber shop. Anyone with access to the data might be able to locate the household, and this violates the household's confidentiality.

Key Concept

Informed consent is a cornerstone in the protection of human subjects. It requires that respondents have a clear understanding of the purpose, procedures, risks, and benefits of the data collection in which they are asked to participate.

Ensuring Reliable and Credible Evaluations through Open Science

One of the fundamental objectives of impact evaluation is to estimate the impact of a program on a range of outcomes of interest. Part 2 discussed a series of methods to ensure that the estimated impacts are robust. A well-designed and well-implemented impact evaluation should ensure that results are unbiased, reliable, and credible, and that they contribute to a wider body of knowledge. When evaluations are unbiased, reliable, and credible, and can be interpreted within a relevant body of related knowledge, they can contribute to good policy decisions and to improving people's lives. In reality, however, several issues can impede the attainment of this ideal.

In this section, we will discuss how a number of scientific issues in impact evaluation can translate into difficult issues for policy makers, and we will discuss potential measures to prevent or mitigate these issues. These measures are commonly grouped under the term *open science*, because they aim to make research methods transparent.[5] Most of these issues need to be handled by the research team, but the policy team overseeing an impact evaluation needs to be aware of them while managing impact evaluations. Issues, policy implications, and possible solutions are summarized in table 13.1.

Table 13.1 Ensuring Reliable and Credible Information for Policy through Open Science

Research issue	Policy implications	Prevention and mitigation solutions through open science
Publication bias. Only positive results are published. Evaluations showing limited or no impacts are not widely disseminated.	Policy decisions are based on a distorted body of knowledge. Policy makers have little information on what *doesn't* work and continue to try out/adopt policies that have no impact.	Trial registries
Data mining. Data are sliced and diced until a positive regression result appears, or the hypothesis is retrofitted to the results.	Policy decisions to adopt interventions may be based on unwarranted positive estimates of impacts.	Preanalysis plans
Multiple hypothesis testing, subgroup analysis. Researchers slice and dice the data until they find a positive result for some group. In particular, (1) multiple testing leads to a conclusion that some impacts exist when they do not, or (2) only the impacts that are significant are reported.	Policy decisions to adopt interventions may be based on unwarranted positive estimates of impacts.	Preanalysis plans and specialized statistical adjustment techniques such as index tests, family-wise error rate, and false discovery rate control[a]
Lack of replication. Results cannot be replicated because the research protocol, data, and analysis methods are not sufficiently documented.	Policy may be based on manipulated (positive or negative) results, as results may be due to mistakes in calculations.	Data documentation and registration, including project protocols, organizing codes, publication of codes, and publication of data
Mistakes and manipulations may go undetected.	Results between different studies cannot be compared.	
Researchers are not interested in replicating studies, and journals are not interested in "me-too" results.	Validity of results in another context cannot be tested.	Changes in journal policies and funding policies to require data documentation and encourage replication
Interventions cannot be replicated because the intervention protocol is not sufficiently documented.	Policy makers may be unable to replicate the intervention in a different context.	

a. For a basic introduction to the multiple comparisons problem and potential statistical corrections, please see https://en.wikipedia.org/wiki/Multiple_comparisons_problem.

Publication Bias and Trial Registries

Researchers who work on impact evaluations normally have an interest in making sure that the results of their evaluations are published in peer-reviewed journals because this helps their own careers. However, most of the results published in journals show positive impacts. This then begs the question of what happens to evaluations that show negative results or that fail to show any significant results. Researchers have almost no incentive to write up nonsignificant results or submit them for publication to peer-reviewed journals because they perceive that there is little interest in the results and that the journals will reject their papers (Franco, Malhotra, and Simonovits 2014). This publication bias is commonly referred to as the "file drawer problem" because results stay in the file drawer and are not disseminated or published. Similar publication bias issues may arise for impact evaluations of specific programs. Policy teams, financers, and governments are more likely to publicize and advertise positive results from a program's evaluation than negative or nonresults. Because of these tendencies, it is difficult to have a clear picture of those interventions that do not work, since the results tend not to be available, and the available body of evidence is rather distorted. Policy makers who try to base their policies on available evidence may not have access to the nonpublished nonresults; as a result, they may continue trying out policies that have been unsuccessful elsewhere.

A partial solution to publication bias is trial registration. Impact evaluation teams should be encouraged to register their trials, and the policy team has an important role to play in ensuring that the research team registers the impact evaluation. Trial registration is very common (and often required) in the medical sciences, but it is just starting to gain ground in the social sciences, including for impact evaluations. Registration implies that the researchers publicly declare their intent to carry out an evaluation before actually doing so, by recording key information about the evaluation in a registry (see box 13.1). As a result, it should be possible to have a complete list of impact evaluations that were carried out, whether the results were positive or not.

Registries are a big step forward in ensuring that the available body of knowledge becomes less distorted. However, many challenges remain. For example, even if it is clear from a registry that an evaluation was carried out, it may not be so easy to obtain information about the results of the evaluation. Impact evaluations may be stopped or may not be well carried out. And even if nonresults from an evaluation are available, these often trigger an additional set of questions that make it difficult to interpret the results: Did the researchers find no results because the evaluation was poorly designed

Box 13.1: Trial Registries for the Social Sciences

Impact evaluations of public policies should normally be registered with social science registries rather than with medical registries, due to the nature of the research. Here are a few examples:

- The American Economic Association's registry for randomized controlled trials can be accessed at http://www.socialscienceregistry.org. As of July 2015, it listed 417 studies in 71 countries.
- The International Initiative for Impact Evaluation (3ie) manages the Registry for International Development Impact Evaluations (RIDIE), which focuses on impact evaluations related to development in low- and middle-income countries. It had registered approximately 64 evaluations as of July 2015.
- The Center for Open Science manages the Open Science Framework (OSF), which has a slightly different focus, but it can also serve as a registry (https://osf.io/). The OSF is a cloud-based management system for research projects, which allows snapshots of research to be created at any point in time, with a persistent URL and time stamp. Researchers can upload their protocol, research hypotheses, data, and code to the OSF and share the resulting web link as a proof of registration.

and carried out, because the program was not well implemented, or because the program truly did not have an impact? As chapter 16 discusses, collecting complementary data through program monitoring or from alternative data sources can help ensure that the results are well interpreted.

Data Mining, Multiple Hypothesis Testing, and Subgroup Analysis

Another potential issue with impact evaluation is *data mining*, the practice of manipulating the data in search of positive results. Data mining can manifest itself in different ways. For example, when data are available, there might be a temptation to run regressions on the data until something positive comes up, and then to retrofit an attractive hypothesis to that result. This is an issue for the following reason: when we run statistical tests for significance of impacts, we need to use a level of significance, say 5 percent. Statistically, 1 in 20 impact tests will come out significant at the 5 percent level, even if the underlying distribution does not warrant an impact (see chapter 15 for a discussion of type I errors). With data mining, one can no longer be sure that an impact result is a genuine result, or whether it comes purely from the statistical properties of the test. This issue is related to the issue of *multiple hypothesis testing*: when a piece of research includes many different hypotheses, there is a high likelihood that at least one of them will be confirmed with a positive test purely by chance (because of the statistical properties of the test), and not because of real impact. A similar

situation arises for subgroup analysis: when the sample is sufficiently large, researchers could try to subdivide it until they find an impact for *some* subgroup. Again, one can no longer be sure that an impact result for that subgroup is a genuine result, or whether it comes purely from the statistical properties of the test.

Another example of data mining is when the decision to continue or stop collecting data is made dependent on an intermediate result: say, a household survey was planned for a sample size of 2,000 households and fieldwork has progressed up to 1,000 households. If this reduced sample produces a positive impact evaluation result and a decision is made to stop the data collection to avoid the risk that additional data might change the results, then this would be data mining. Other examples are excluding certain inconvenient observations or groups, or selectively hiding results that do not fit. While there is no reason to believe that these practices are widespread, just a few high-profile, egregious cases have the potential of undermining impact evaluation as a science. In addition, even lesser cases of data mining have the potential to distort the body of evidence used by policy makers to decide what interventions to start, continue, or discontinue.

A common recommendation to avoid data mining is to use a *preanalysis plan*. Such a plan outlines the analysis methods before the impact evaluation analysis is carried out, thereby clarifying the focus of the evaluation and reducing the potential to alter the methods once the analysis has started. The preanalysis plan should specify the outcomes to be measured, the variables to be constructed and used, the subgroups for which analysis will be conducted, and the basic analytical approaches to be used in estimating impacts. Preanalysis plans should also include the researchers' proposed corrections for multiple hypothesis testing and subgroup testing, if required. For example, testing the impact of an education intervention on six different test scores (math, English, geography, history, science, French) for five different school groups (grades 1 through 5) and two genders (male and female) would yield 60 different hypotheses, one or several of which are bound to have a significant test just by chance. Instead, the researcher could propose to compute one or more indexes that group the indicators together, so as to reduce the number of hypotheses and subgroups.[6]

While a preanalysis plan might help alleviate the concern of data mining, there is also a concern that it might remove some needed flexibility in the kind of analysis carried out by researchers. For example, the preanalysis plan may specify the anticipated channels of impact of an intervention throughout the results chain. However, once the intervention is actually implemented, a whole host of additional, unanticipated factors may suddenly appear. For example, if a government is thinking of

implementing a new way of paying health care providers, one might be able to come up with the possible channels of impact. However, it would be very difficult to anticipate every possible effect that this could have. In some cases, qualitative interviews with providers would be needed to understand exactly how they adapt to the changes and how this is affecting performance. It would be very difficult to incorporate all these possibilities into the preanalysis plan in advance. In that case, researchers would have to work outside of the original preanalysis plan—and should not be penalized for this. In other words, a preanalysis plan can lend additional credibility to evaluations by turning them into confirmations of a hypothesis, rather than just exploratory research; but researchers should be able to continue to explore new options that can be turned into confirmatory research in subsequent evaluations.

Lack of Replication

There are two kinds of replication that are important for impact evaluation. First, for a given study, researchers other than the original research team should be able to produce the same (or at least very similar) results as the original researchers when using the same data and analysis. Replications of a given impact evaluation result are a way to check their internal validity and unbiasedness. When studies or results cannot be replicated because of lack of availability of information about coding or data, there is a risk that mistakes and manipulations in the analysis may go undetected, and that inaccurate results may continue to influence policy. Fortunately, substantial advances are being made in terms of making data, coding, and protocols available. An increasing number of social science journals are starting to require that data and coding be made available along with publication of results. Guidelines such as the Transparency and Openness Promotion Guidelines developed by the Center for Open Science are slowly changing practices and incentives. To ensure that replication can take place, impact evaluation teams need to make data publicly available and ensure that all protocols (including the randomization protocol), data sets, and analysis codes of the impact evaluation are documented, safely stored, and sufficiently detailed.

Second, once an evaluation is completed, it should be possible for other policy makers and researchers to take the original intervention and evaluation protocols and apply them in a different context or at a different time to see if the results hold under different circumstances. Lack of replication of evaluation results is a serious issue for policy makers. Say an evaluation shows that introducing computers in schools has highly beneficial results, but this is the only study that produced such results, and other

researchers are unable to get the same positive results in subsequent evaluations of similar programs. What is a policy maker to do in such cases? Lack of replication of results can have many causes. First, it can be difficult to carry out evaluations that just try to replicate results that were obtained in a previous study: neither researchers nor financers might be interested in "me-too" studies. Second, even when there are willingness and funds to replicate studies, replication is not always possible because the protocols (including the randomization protocol), data, and analysis code of the original study might not be available or sufficiently detailed. There is a growing effort among organizations that support impact evaluations to encourage replications across settings, for instance, by developing clusters of studies on similar topics or fostering multisite impact evaluations.

Checklist: An Ethical and Credible Impact Evaluation

Policy makers have an important role to play in ensuring that the right stage is set for an ethical and credible impact evaluation. In particular, policy makers bear the primary responsibility for ensuring that the program assignment rules are fair, and they should hold the research team accountable for the transparency of the research methods. We suggest the following checklist of questions to ask:

- ✓ Is assignment to the treatment and comparison groups fair? Are there any groups with particularly high need that should receive the program in any case? Who will be excluded from the impact evaluation?
- ✓ Has the research team identified the relevant Institutional Review Board or National Ethics Review Committee?
- ✓ Does the impact evaluation schedule allow sufficient time to prepare and submit the research protocol to the IRB and obtain consent before data collection from human subjects begins?
- ✓ Did the research team submit the research protocol and preanalysis plan to a social science trial registry?
- ✓ Is a procedure in place to ensure that the key elements of the intervention are documented as they happen, and not only as they are planned?
- ✓ Do policy makers understand that evaluation results might show that the intervention was not effective, and do they agree that such results will be published and not held back?

✓ Has the evaluation team identified the way in which evaluation data and results will be made available, even if the research team does not manage to publish the results in a peer-reviewed journal?

The principles, issues, and checklist identified in this chapter can help ensure that your impact evaluation is both credible and ethical.

Additional Resources

- For accompanying material to the book and hyperlinks to additional resources, please see the Impact Evaluation in Practice website (http://www.worldbank.org/ieinpractice).
- Human Subjects training from the U.S. National Institutes of Health (NIH)
  - The NIH offers an online training that—while focused on medical sciences and the United States—is still very informative and takes only about one hour to complete. See http://phrp.nihtraining.com/users/login.php and http://www.ohsr.od.nih.gov.
- Human Subjects training through the Collaborative Institutional Training Initiative at the University of Miami (CITI)
  - CITI offers international courses in several languages to both organizations and individuals, though the program has a fee (starting at US$100 per person). See http://www.citiprogram.com.
- International compilation of human research standards
  - Every year, the U.S. Department of Health and Human Services publishes a compilation of laws, regulations, and guidelines that govern research involving human subjects. The 2015 edition includes 113 countries, as well as the standards from a number of international and regional organizations. The document identifies national and international institutional review boards (http://www.hhs.gov/ohrp/international).
- Procedures for Protection of Human Subjects in Research Supported by USAID (U.S. Agency for International Development) (http://www.usaid.gov/policy/ads/200/humansub.pdf).
- Manual of Best Practices in Transparent Social Science Research, by Garret Christensen with assistance from Courtney Soderberg (Center for Open Science) (https://github.com/garretchristensen/BestPracticesManual).
  - This is a working guide to the latest best practices for transparent quantitative social science research. The manual is regularly updated.
- The Transparency and Openness Promotion (TOP) Guidelines (http://centerforopenscience.org/top/).
  - The guidelines can be found on the website of the Center for Open Science.
- For links to recognized independent review boards and independent IRB services, see the Inter-American Development Bank Evaluation Portal (http://www.iadb.org/evaluationhub).
- For more on data collection, see the Inter-American Development Bank Evaluation Portal (http://www.iadb.org/evaluationhub).

  - See the data collection section under Protection of Human Subjects.
  - Note the link to the Association for the Accreditation of Human Research Protection Programs (AAHRPP). AAHRPP provides training and certification for IRBs. A list of accredited organizations can be found on their website.
- For guidelines for protecting human research participants, see the World Bank Impact Evaluation Toolkit, Module 4 (http://www.worldbank.org/health/impactevaluationtoolkit).

Notes

1. In the absence of national ethical guidelines, the investigator and team should be guided by the Helsinki Declaration adopted by the Twenty-Ninth World Medical Assembly in Tokyo (October 1975) and Article 7 of the International Covenant of Civil and Political Rights, adopted by the United Nations General Assembly on December 16, 1966. Additional guidance is provided by the World Health Organization and by the *Belmont Report on Ethical Principles and Guidelines for the Protection of Human Subjects* (1974) (http://www.hhs.gov/ohrp/policy/belmont.html). An international compilation of human research standards can be found at http://www.hhs.gov/ohrp/international.
2. The World Health Organization's guidelines on how to write a protocol for research involving human participation can be found at http://www.who.int/rpc/research_ethics/guide_rp/en/index.html.
3. More information on consent procedures during data collection can be found in the World Bank's Impact Evaluation Toolkit.
4. More information on the assignment of IDs can be found in the World Bank's Impact Evaluation Toolkit.
5. For more information on open science recommendations in the context of impact evaluation, please see Miguel and others (2014).
6. Other techniques are available. See, for example, Anderson (2008).

References

Anderson, Michael L. 2008. "Multiple Inference and Gender Differences in the Effects of Early Intervention: A Reevaluation of the Abecedarian, Perry Preschool, and Early Training Projects." *Journal of the American Statistical Association* 103 (484): 1481–95.

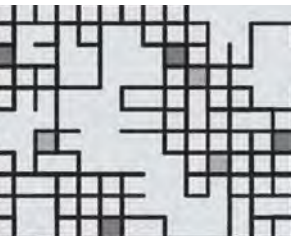
Christensen, Garret, with Courtney Soderberg. 2015. *The Research Transparency Manual*. Berkeley Initiative for Transparency in the Social Sciences. https://github.com/garretchristensen/BestPracticesManual.

Franco, Annie, Neil Malhotra, and Gabor Simonovits. 2014. "Publication Bias in the Social Sciences: Unlocking the File Drawer." *Science* 345 (6203): 1502–5.

Miguel, Edward, C. Camerer, Katherine Casey, Joshua Cohen, Kevin M. Esterling, and others. 2014. "Promoting Transparency in Social Science Research." *Science* 343: 30–31.

National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. 1978. *The Belmont Report: Ethical Principles and Guidelines for the Protection of Human Subjects of Research.* U. S. Department of Health, Education, and Welfare Publication No. (OS) 78-0012. Washington, DC: Government Printing Office.

Vermeersch, Christel, Elisa Rothenbühler, and Jennifer Sturdy. 2012. *Impact Evaluation Toolkit: Measuring the Impact of Results-Based Financing on Maternal and Child Health*. World Bank, Washington, DC. http://www.worldbank.org/health/impactevaluationtoolkit.



CHAPTER 14

Disseminating Results and Achieving Policy Impact

A Solid Evidence Base for Policy

You have finally completed the arduous task of evaluating your program from start to finish, a multiyear effort that involved significant financial and human resources. The final evaluation products, including a 200-page report, complete with multiple annexes, have been delivered. Mission accomplished?

Actually, now a new phase begins to ensure that all this effort pays off in the form of policy impact. Impact evaluations fundamentally aim to provide accountability for past investments and guide policy decisions in the future toward more cost-effective development so that scarce resources yield the highest social returns possible. Those policy decisions will be influenced by a range of factors, from the political economy to ideological positions. But impact evaluations can and should influence policy by providing a solid evidence base that guides resources toward effective, proven interventions. From the earliest stages of a new program, even while it is being conceived, evidence from available relevant impact evaluations should play a central role in informing the program's design and guiding the next set of evaluation questions.

Key Concept

Impact evaluations must answer relevant policy questions in a rigorous manner, bring actionable evidence to key stakeholders in a timely manner, and disseminate evidence in a form that decision makers can easily access and use.

Typically, however, the process of influencing policy does not happen spontaneously through the generation of evidence alone. Impact evaluations must first and foremost answer relevant policy questions in a rigorous manner, bringing actionable evidence to key stakeholders in a timely manner. But policy makers and program managers may not have the time and energy to delve into the details of a 200-page report, trying to distill the key findings and recommendations. Information generated through impact evaluations needs to be packaged and disseminated in a way that decision makers can easily access and use.

In this chapter, we discuss ways your impact evaluation can influence policy, key constituencies you may want to reach, and strategies for communicating and disseminating information to target audiences so that the evaluation achieves policy impact.

The starting point for influencing policy is the selection of relevant evaluation questions that will be useful for making policy decisions, as discussed in part 1 of this book. During the very earliest stages of designing an impact evaluation, policy makers and evaluators will likely start with a wish list of questions. These questions should be vetted with the key group of stakeholders and decision makers who will ultimately use the impact evaluation to formulate decisions. The wish list will typically be adjusted and improved over time to include a more limited number of well-formulated questions that are both policy relevant and amenable to being answered through an impact evaluation, using the methods discussed in part 2 of this book. Simultaneously engaging policy makers to identify the important questions and the evaluation team to gauge the technical feasibility of answering those questions is a critical first step to influencing policy.

Once the program is up and running, the impact evaluation will probably produce important analytical inputs that can serve to inform policy well before the program and impact evaluation have come to fruition. One common example is the findings of a baseline survey or an analysis of short-term results. Baseline surveys often produce the first comprehensive and population-specific data for a program, providing descriptive statistics that can be fed into the program design and policy dialogue. While a program may have a general description of its target population through national surveys or diagnostic studies, the baseline survey may provide the first detailed information for specific subpopulations or geographic areas where the program will operate. For example, a program designed to improve child nutrition through nutritional supplementation may have statistics on rates of stunting and wasting at a national level from existing surveys, but the baseline survey might provide the first measures of nutritional status and eating habits for the group of children that the program will actually serve. This type of information can be valuable for tailoring the intervention design and must

be made available to the policy team in a timely manner (ideally before the intervention is rolled out) in order to influence the program's design. Box 14.1 presents an example from Mozambique.

Some impact evaluations, particularly those that rely on administrative data sources or routine surveys, can produce intermediate results that feed back to the program while the program is being implemented. These results provide valuable information and recommendations on how indicators along the causal pathway are changing over time, allowing both the implementation of the program and timing of evaluation activities to be adjusted accordingly. For example, if half way through a program, it is clear that there are no effects on short-term outcomes, the program may be advised to implement an operational evaluation to detect bottlenecks and undertake corrective actions. The evaluation timeline could be adjusted to avoid conducting a costly endline survey before the results of the intervention have had a chance to kick in. In the child nutrition example, if the analysis of administrative data on the distribution of nutritional supplements shows that supplements are not reaching the intended beneficiaries, then the

Box 14.1: The Policy Impact of an Innovative Preschool Model in Mozambique

(continued from chapter 1)

Recall that in chapter 1 (box 1.2), an evaluation of Save the Children's community-based preschool program in Mozambique was an important input for the country's national early childhood development policy. However, even before the program ended, the evaluation generated new and revealing information for the country's policy debate in this area. The evaluation's baseline survey generated the first population-based measurements of child development outcomes, using specialized tests of child development adapted to the Mozambican context, and collected by specialized surveyors. Even though data were from a select group of communities in one province of Mozambique, the baseline statistics provided a first snapshot of child development outcomes in the country, showing that many children lagged behind in a number of dimensions, from language and communication to cognitive and socioemotional development.

The baseline survey was presented by the evaluation team in seminars and workshops, where results were discussed with high-level policy makers, international donors, and key stakeholders from the early childhood development community. The data generated through the impact evaluation further bolstered the need for investing in this area, and played a catalytic role in mobilizing support for the early childhood agenda in the country. The completed evaluation was eventually disseminated through various outlets, including policy notes, videos, and blogs, a number of which have been compiled on the website of the International Initiative for Impact Evaluation (3ie).

policy team can be alerted that a review of its supply chain is in order. The costly follow-up survey for measuring child height and weight could be postponed until some months after the program is operating effectively, since there is no good reason to believe that the nutritional program will generate impacts any sooner if it was not reaching its participants.

Impact evaluations tend to produce large volumes of information, from the technical underpinnings of the evaluation design, to descriptive statistics and impact analyses complete with data sets, statistical code, and reports. It is critical that the evaluation team make an effort to document all information throughout the evaluation cycle, and to the extent possible, put relevant (nonconfidential) technical documentation in the public domain: for example, through a dedicated website. Ultimately, the credibility of the evaluation results will hinge on the methodology and rigor with which the evaluation was implemented. Full transparency strengthens the trustworthiness of the evaluation and its potential for influencing policy.

While completeness and transparency are critical, most consumers of the information will not delve into the details. It will be up to the evaluation team to distill a manageable set of key messages summarizing the most policy-relevant results and recommendations, and to communicate these messages consistently across audiences. The sequencing of dissemination activities is also critical for policy impact. Unless otherwise agreed on by the policy team, the initial round of presentations and consultations of an evaluation's results should be conducted internally, with program staff, managers, and policy makers. A premature result, leaked to the public domain, can hurt a program's reputation, with lasting harm for the evaluation's policy impact.

Tailoring a Communication Strategy to Different Audiences

There are at least three primary audiences for impact evaluation findings: program staff and managers involved in the specific program being evaluated; high-level policy makers who will use the evaluation to inform funding and policy design decisions; and the community of practice, broadly encompassing the academic community, development practitioners, civil society (including the media), and program participants. Each of these audiences will have different interests in the evaluation results and will require tailored communication strategies in order to accomplish the objective of informing and influencing policy (table 14.1).

Technicians and managers. The first key audience includes technical and operational staff, and managers who designed and implemented the

Table 14.1 Engaging Key Constituencies for Policy Impact: Why, When, and How

	Program staff and managers	High-level policy makers	Development academics and civil society groups
Why?	They can become champions of impact evaluation and the use of evidence.	They need to understand why the issue is important, how impact evaluation can help them make better decisions, and ultimately, what the evidence tells them about where their energies (and available financing) should be directed.	They need evidence about the impact of development programs in order to make decisions, design new programs, replicate successful programs in other countries, and carry out research that can help improve lives.
When?	Early on, even before the program is rolled out, and with continued and frequent interactions throughout. Baseline data can be used to tailor the intervention. They are the first to comment on evaluation results.	Early on, when defining the evaluation questions and before the evaluation begins, and again when results have been finalized. It's important that senior policy makers understand why an impact evaluation is being conducted and how the results can help them.	Depending on the program being evaluated, civil society groups and development experts can be important local champions. Information should be disseminated once results are finalized and have been vetted by program staff and policy makers.
How?	Introduce the role of evidence in policy making in a workshop to engage program managers in the evaluation design. Follow up with meetings at key points: immediately after collection of baseline data, after collection of intermediate results, and at the endline.	Present at national workshops and seek direct meetings with senior-level staff to explain the work. Encourage program managers, technical staff, and mid-level policy makers to keep ministries informed about the impact evaluation. When the evidence is finalized, present to senior policy makers. When possible, include cost-benefit or cost-effectiveness analysis and suggestions for next steps.	Public events and forums—including seminars and conferences, working papers, journal articles, media coverage, and web-based materials—are all avenues for reaching these audiences.

program, as well as individuals from institutions (such as a ministry or funding institution) closely associated with the project. This group of individuals will typically be the first to see the evaluation results and provide comments on the evaluation's interpretations and recommendations.

Since this is the first time results usually see the light of day, timing the release of information to this key constituency is critical. On the one hand, it is important to share the results early on, so program decision makers can incorporate changes and make policy decisions, such as scaling the intervention up (or down) or adjusting program components to improve the use of resources and achieve greater impact. On the other hand, we caution against sharing very preliminary results based on partial or

incomplete analysis. These results could be subject to change. Their release could set expectations with program staff and prompt premature policy decisions that could be costly to reverse in the future. Thus an appropriate balance of timeliness and completeness should be sought for the initial dissemination of results with the project team. This typically happens when the evaluation team has conducted a thorough analysis and robustness checks, but before the final results, interpretation, and recommendations are formulated.

The program staff and managers will usually be interested in both the technical details of the evaluation methodology and analysis and the particulars of the initial findings and recommendations. The initial discussions of results with this group may be well suited for workshop-style meetings, with presentations by the evaluation team and ample time for clarifying questions and comments from all sides. These initial discussions will typically enrich the final analysis, inform the interpretation of results, and help tailor the final recommendations so they are best suited to guide the program's policy objectives. The initial discussions with program staff and managers will be a good opportunity to discuss unexpected or potentially controversial results, and to propose policy recommendations and responses in anticipation of public disclosure of the impact evaluation.

Negative results (including finding no impact) or unexpected results can be disappointing for program staff and managers who have invested significant time and energy into a program, but they also serve the critical function of prompting policy to be reformulated. For example, if the program is found to have failed to achieve its primary objective because of implementation challenges, measures can be taken to address those areas and an improved program can be reevaluated later. If the program does not produce impacts in the short term or only produces impacts in a subset of the results chain, and there is reason to believe that additional time is required to reach final outcomes, then the evaluation can present and defend the initial results, and additional measurements can be planned at a future date. Finally, if it is clear that the intervention is failing to generate its intended benefits or is unexpectedly causing harm, then the program managers can take immediate steps to stop the intervention or reformulate its design. In this way, when the evaluation results are made public, policy makers in charge of the program can announce corrective measures and formulate responses ahead of time, in anticipation of tough questions in policy debates or the media.

High-level policy makers. The second key constituency is high-level policy makers who will make policy decisions based on the results of the impact evaluation, such as whether to expand, maintain, or decrease funding for an intervention. These may include the national legislature, presidents and prime ministers, ministers and principal secretaries, board of directors,

or donors. This group of stakeholders will typically be provided with the evaluation results once they are finalized and have been reviewed by program staff and managers and vetted by external technical experts. At this stage, the evaluation team will need to focus on communicating the key results and recommendations in an accessible manner; technical details of the evaluation may be of secondary importance. High-level policy makers will be interested in the translation of impacts into economically meaningful values through cost-benefit analysis, or comparison with alternative interventions through cost-effectiveness analysis. These parameters will help inform decision makers as to whether the program is a worthwhile way to invest limited resources to further an important development objective. High-level policy makers may also be interested in using the results to further their political agenda, such as lobbying for (or against) a given public policy that the evaluation does (or does not) support. The evaluation team can collaborate with communication experts to ensure that the results and related recommendations are correctly interpreted and that messages in the communications strategy remain aligned with the evaluation findings.

The community of practice. The third key constituency for achieving a policy impact broadly encompasses the consumers of evaluation outside the direct realm of the program or country context. This heterogeneous group encompasses the community of practice in sectors germane to the evaluation, including development practitioners, academia, civil society, and policy makers in other countries. Development practitioners beyond the specific program may be interested in using the results of the evaluation to inform the design of new or existing programs. These practitioners will be interested both in details of the evaluation (methods, results, recommendations) and in operational lessons and recommendations that can help implement their own projects more effectively. The academic community, on the other hand, may be more interested in the evaluation's methodology, data, and empirical findings.

Within civil society at large, two key constituencies stand out: the media and program participants. Informing the public of the results of an evaluation through the media can play a key role in achieving accountability for public spending, building public support for the evaluation recommendations, and sustaining effective policies. This is particularly true of new and innovative policies where the outcome was initially uncertain or the subject of controversy in the policy debate. If the evaluation sheds empirical light on what had been to date a largely theoretical or ideological debate, it can be a powerful instrument for policy change.

Finally, program participants should be included in the dissemination efforts. Participants have invested their time and energy in the program and may have spent considerable time providing information for purposes of

the evaluation. Ensuring that program participants have access to and remain informed about the evaluation results is a small but significant gesture that can contribute to their continued interest in the program and willingness to participate in future evaluations.

Disseminating Results

Next, we discuss a variety of strategies that can be considered to inform these key constituencies and achieve policy impact. Ideally, the early stages of the evaluation planning will include a dissemination or policy impact strategy. This strategy should be agreed to up front, clearly spelling out the evaluation policy objective (for example, expansion of a more cost-effective intervention model), the key audiences that the evaluation intends to reach, the communication strategies to be used, and a budget for conducting dissemination activities. While the format and content of the dissemination activities and products will vary on a case-by-case basis, we provide some tips and general guidelines in the remainder of this chapter. Box 14.2 lists some outreach and dissemination tools.

Reports are typically the first outlet for the complete set of evaluation results. We recommend keeping reports to a moderate length, in the range of 30 to 50 pages, including an abstract of 1 page or less, and a 2 to 4-page executive summary with the principal results and recommendations. Technical details, associated documentation, and supporting analysis such as robustness and falsification tests can be presented in annexes or appendixes.

Box 14.2: Outreach and Dissemination Tools

Here are some examples of outlets for disseminating impact evaluations:

- Slide shows about the program and evaluation results
- Videos that feature beneficiaries giving their view of the program and how it affects their lives
- Short policy notes explaining the evaluation and summarizing policy recommendations
- Blogs by researchers and policy makers that explain the importance of the evaluation
- Full reports after final results have come in, with strong executive summaries to ensure that readers can quickly understand the main findings
- Media invitations that let journalists see the program in action and report results.

Publishing the impact evaluation as an academic working paper or article in a peer-reviewed scientific journal can be a laborious but very worthwhile final step in writing up the evaluation results. The rigorous peer reviews required for the publication process will provide valuable feedback for improving the analysis and interpretation of results, and publication can provide a strong signal to policy makers as to the quality and credibility of an evaluation's results.

Based on the agreed dissemination strategy, reports and papers can be published through various outlets, including on the program website; through the evaluating institution's website; and as part of working paper series, peer-reviewed academic journals, and books.

While evaluation reports and academic papers serve as the foundation for the dissemination strategy, their reach to a broader audience outside the community of practice and academia may be limited by their length and technical language. The evaluation team, perhaps in collaboration with communication experts, might find it useful to produce short articles written in a storytelling or journalistic fashion, with clear and simple language for dissemination to broader audiences. Short articles can be published in the form of policy briefs, newsletters, bulletins, and infographics. For these publications it will be particularly helpful to eliminate technical jargon and translate results into visually appealing representations, including pictures, charts, and graphs (box 14.3).

Box 14.3: Disseminating Impact Evaluations Effectively

Various publications showcase the results of impact evaluations in an accessible and user-friendly format. These include two updates with a regional focus:

- Impact evaluation results from programs throughout Latin America and the Caribbean are featured in the Development Effectiveness Overview, published yearly by the Office of Strategic Planning and Development Effectiveness at the Inter-American Development Bank (IDB). Results are summarized in short, easy-to-read articles, which include one-page infographic summaries that distill the key impact evaluation question, methods, results, and policy recommendations using figures, graphics, and icons that allow readers to grasp the key messages very quickly and intuitively. The 2014 Development Effectiveness Overview includes results from impact evaluations of programs as diverse as tourism in Argentina, job training in the Dominican Republic, agricultural productivity in Bolivia, and youth orchestras in Peru.
- The World Bank's Africa Impact Evaluation Update brings together the latest evidence from the region. It focused on gender in 2013 and on agriculture and land in 2014.

Sources: http://deo.iadb.org and http://www.worldbank.org.

Evaluation teams can generate a set of presentations that accompany written reports and short articles. Presentations should be tailored to the specific audience. A good starting point is to produce a technical presentation for project staff and academic audiences, and another shorter and less technical presentation for policy makers and civil society. While the key findings and policy recommendations will be the same, the structure and content of these two types of presentation will have important variations. The technical presentation should focus on building credibility for the results through presentation of the evaluation methods, data, and analysis before reaching results and recommendations. A presentation targeted to policy makers should emphasize the development problem that the intervention is meant to address and the practical implications of findings, while skimming over technical details.

To take advantage of expanding access to the Internet in developing countries and low-cost alternatives for producing multimedia, evaluation teams can also consider a range of media to disseminate evaluation findings, from websites to audio and video pieces. Short video clips can be a powerful way to transmit complex ideas through images and sound, allowing the evaluation story to unfold in a way that is more quickly and fully comprehensible than typical print media (box 14.4).

Finally, armed with a variety of dissemination products, the evaluation team must be proactive about disseminating these products to consumers within the program, government, and broader community of practice so they reach the intended users and can be assimilated into the decision-making process and policy debate. The process of dissemination happens

Box 14.4: Disseminating Impact Evaluations Online

Here are some noteworthy examples of online dissemination of impact evaluation results:

- The International Initiative for Impact Evaluation (3ie) organizes evidence from impact evaluations by sector, including policy briefs, systematic reviews, and evidence gap maps.
- The Abdul Latif Jameel Poverty Action Lab (J-PAL) disseminates evidence from impact evaluations conducted by affiliated researchers, including policy briefs, cost-effectiveness analysis and links to academic papers.
- The World Bank's Development Impact Evaluation (DIME) presents briefs, newsletters, and reports highlighting results from impact evaluations of World Bank projects.
- The World Bank's Strategic Impact Evaluation Fund (SIEF) includes videos, briefs, and interviews.

Box 14.5: Impact Evaluation Blogs

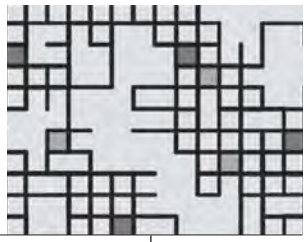
Here are a few examples of blogs that regularly feature the results of impact evaluations:

- World Bank Development Impact Blog
- Inter-American Bank Development Effectiveness Blog
- Innovations for Poverty Action Blog.

through face-to-face meetings between the evaluation team and program manager, lobbying with high-level policy makers, presentations in seminars and conferences where academics and members of the community of practice gather to learn about the latest developments in development research and evaluation, interviews and news programs on radio and television, and increasingly through the Internet. Blogs and social media in particular can be cost-effective ways to reach large numbers of potential users and to capture traffic and guide readers toward the array of products available from a given evaluation (box 14.5). While the particular strategies will vary on a case-by-case basis, we again recommend planning and budgeting the dissemination outlets and activities early on, so that the results of the evaluation can reach their intended audiences quickly and effectively, thus maximizing the policy impact.

Additional Resources

- For accompanying material to the book and hyperlinks to additional resources, please see the *Impact Evaluation in Practice* website (http://www.worldbank.org/ieinpractice).
- The International Initiative for Impact Evaluation (3ie) and the Overseas Development Institute (ODI) have developed an online Policy Impact Toolkit to help disseminate and use evidence from impact evaluations for decision making.



Part 4

HOW TO GET DATA FOR AN IMPACT EVALUATION

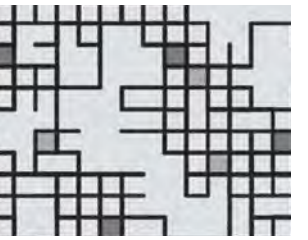
The fourth part of the book provides guidance on how to get data for an impact evaluation, including choosing the sample and finding adequate sources of data.

Chapter 15 discusses how to draw a sample from a population of interest, and how to conduct power calculations to determine the appropriate size of the impact evaluation sample. The chapter focuses on describing the main intuition behind sampling and power calculations. It also highlights the elements that the policy team needs to provide to the research team or technical expert responsible for undertaking sampling and power calculations.

Chapter 16 reviews the various sources of data that impact evaluations can use. It highlights when existing sources of data can be used, including administrative data. Since many evaluations require the collection of new data,

the chapter discusses the steps in collecting new survey data: determining who will collect the data, developing and piloting data collection instruments, conducting fieldwork and quality control, and processing and storing data.

Chapter 17 provides a conclusion to the overall book. It briefly reviews the core elements of a well-designed impact evaluation, as well as some tips to mitigate common risks in conducting an impact evaluation. It also provides some perspectives on recent growth in the use of impact evaluation and related institutionalization efforts.



CHAPTER 15

Choosing a Sample

Sampling and Power Calculations

Once you have chosen a method to select a comparison group and estimate the counterfactual, one of the next steps in undertaking an impact evaluation is to determine what data you will need and the sample required to precisely estimate differences in outcomes between the treatment group and the comparison group. In this chapter, we discuss how you can draw a sample from a population of interest (sampling) and how you can determine how large the sample needs to be to provide precise estimates of program impact (power calculations). Sampling and power calculations require specific technical skills and are often commissioned to a dedicated expert. In this chapter, we describe the basics of performing sampling and power calculations, and we highlight the elements that the policy team needs to be able to provide to technical experts.

Drawing a Sample

Sampling is the process of drawing units from a population of interest to estimate the characteristics of that population. Sampling is often necessary, as typically it is not possible to directly observe and measure outcomes for the entire population of interest. For instance, if you are interested in knowing the average height of children below age two in a country, it would be very

hard, costly, and time consuming to directly visit and measure all children in the population. Instead, a sample of children drawn from the population can be used to infer the average characteristics in the population (figure 15.1).

The process by which a sample is drawn from the population of interest is crucial. The principles of sampling provide guidance to draw representative samples. In practice, there are three main steps to draw a sample:

1. Determine the population of interest.
2. Identify a sampling frame.
3. Draw as many units from the sampling frame as required by power calculations.

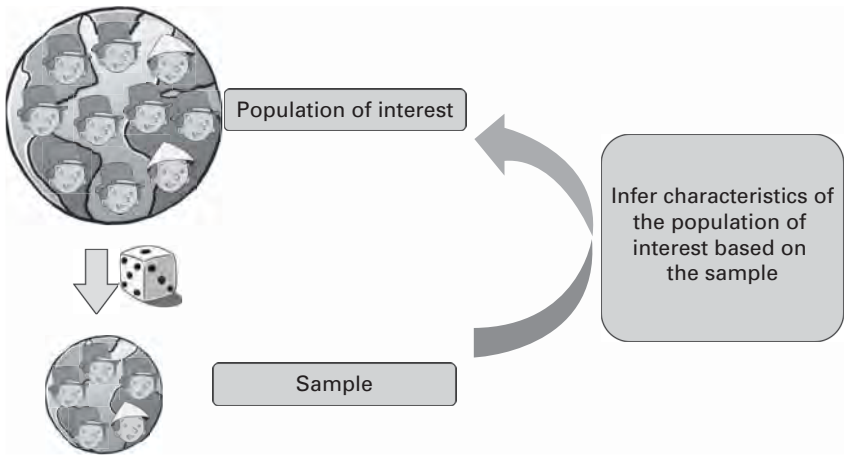
First, the *population of interest* needs to be clearly defined. This requires accurately specifying the unit within the population of interest for which outcomes will be measured, and clearly defining the geographic coverage or any other relevant attributes that characterize the population of interest. For example, if you are managing an early childhood development program, you may be interested in measuring the impact of the program on cognitive outcomes for young children between ages three and six in the entire country, only for children in rural areas, or only for children enrolled in preschool.

Key Concept

A sampling frame is the most comprehensive list that can be obtained of units in the population of interest. A coverage bias occurs if the sampling frame does not perfectly overlap with the population of interest.

Second, once the population of interest has been defined, a *sampling frame* must be established. The sampling frame is the most comprehensive list that can be obtained of units in the population of interest. Ideally, the

Figure 15.1 Using a Sample to Infer Average Characteristics of the Population of Interest

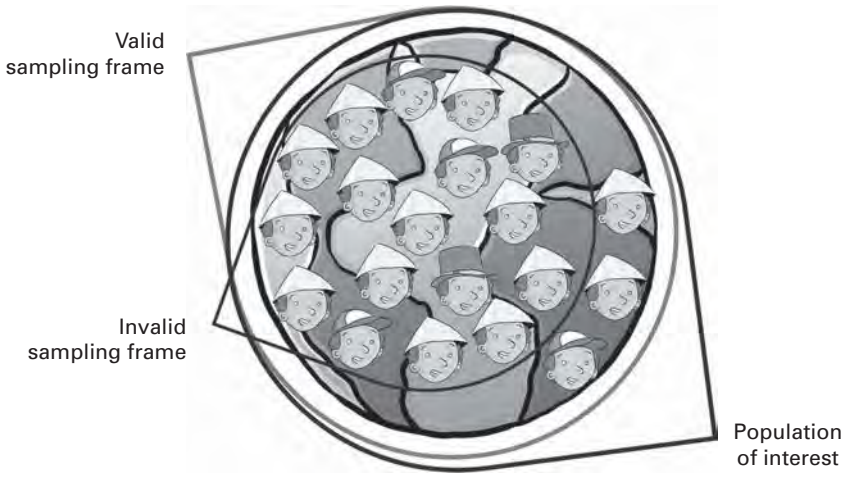


sampling frame should exactly coincide with the population of interest. A full and totally up-to-date census of the population of interest would constitute an ideal sampling frame. In practice, existing lists, such as population censuses, facility censuses, or enrollment listings, are often used as sampling frames.

An adequate sampling frame is required to ensure that the conclusions reached from analyzing a sample can be generalized to the entire population. Indeed, a sampling frame that does not exactly coincide with the population of interest creates a *coverage bias*, as illustrated in figure 15.2. If coverage bias occurs, results from the sample do not have external validity for the entire population of interest, but only for the population included in the sampling frame. The degree to which statistics computed from the sample can be generalized to the population of interest as a whole depends on the magnitude of the coverage bias, in other words, the lack of overlap between the sampling frame and the population of interest.

Coverage biases constitute a risk, and the construction of sampling frames requires careful effort. For instance, census data may contain the list of all units in a population. However, if much time has elapsed between the census and the time the sample data are collected, the sampling frame may no longer be fully up to date. Moreover, census data may not contain sufficient information on specific attributes to build a sampling frame. If the population of interest consists of children attending preschool, and the census does not contain data on preschool enrollment, complementary enrollment data or facility listings would be needed.

Figure 15.2 A Valid Sampling Frame Covers the Entire Population of Interest



Key Concept

Sampling is the process by which units are drawn from a sampling frame. Probabilistic sampling assigns a well-defined probability for each unit to be drawn.

Once you have identified the population of interest and a sampling frame, you must choose a method to draw the sample. Various alternative procedures can be used.

Probabilistic sampling methods are the most rigorous, as they assign a well-defined probability for each unit to be drawn. The three main probabilistic sampling methods are the following:

- *Random sampling.* Every unit in the population has exactly the same probability of being drawn.[1]
- *Stratified random sampling.* The population is divided into groups (for example, male and female), and random sampling is performed within each group. As a result, every unit in each group (or stratum) has the same probability of being drawn. Provided that each group is large enough, stratified sampling makes it possible to draw inferences about outcomes not only at the level of the population but also within each group. Stratified sampling is useful when you would like to oversample subgroups in the population that are small (like minorities) in order to study them more carefully. Stratification is essential for evaluations that aim to compare program impacts between such subgroups.
- *Cluster sampling.* Units are grouped in clusters, and a random sample of clusters is drawn. Thereafter, either all units in those clusters constitute the sample or a number of units within the cluster are randomly drawn. This means that each cluster has a well-defined probability of being selected, and units within a selected cluster also have a well-defined probability of being drawn.

In the context of an impact evaluation, the procedure for drawing a sample is often determined by the eligibility rules of the program under evaluation. As will be described in the discussion on sample size, if the smallest viable unit of implementation is larger than the unit of observation, randomized assignment of benefits will create clusters. For this reason, cluster sampling often arises in impact evaluation studies.

Nonprobabilistic sampling can create serious sampling errors. For instance, suppose that a national survey is undertaken by asking a group of interviewers to collect household data from the dwelling closest to the school in each village. When such a nonprobabilistic sampling procedure is used, it is likely that the sample will not be representative of the population of interest as a whole. In particular, a coverage bias will arise, as remote dwellings will not be surveyed.

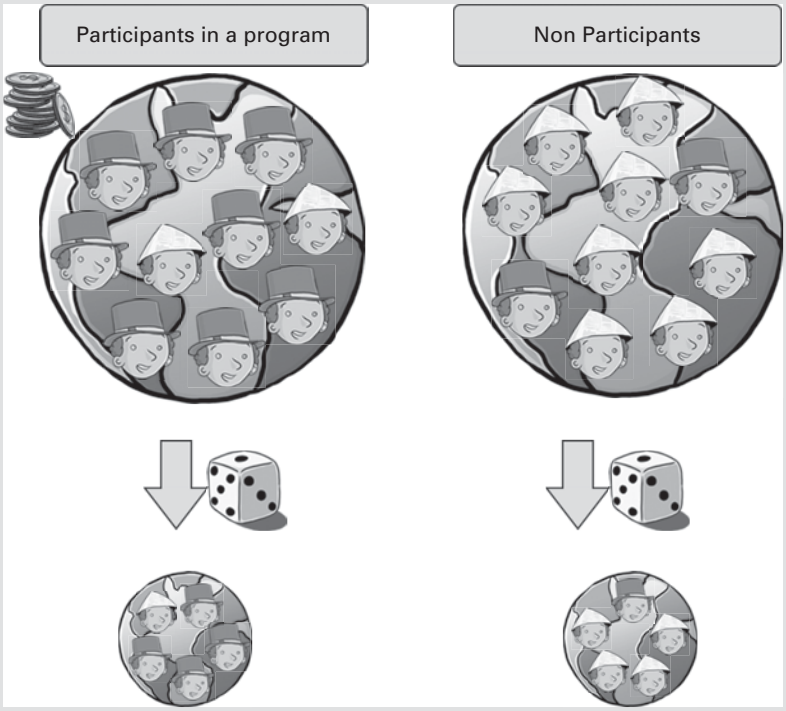
It is necessary to pay careful attention to the sampling frame and the sampling procedure to determine whether results obtained from a given sample can be generalized to the entire population of interest. Even if the

sampling frame has perfect coverage and a probabilistic sampling procedure is used, nonsampling errors can also affect the internal and external validity of the impact evaluation. Nonsampling errors are discussed in chapter 16. Lastly, there is sometimes confusion between random sampling and randomized assignment. Box 15.1 makes clear that random sampling is very different from randomized assignment.

Box 15.1: Random Sampling Is Not Sufficient for Impact Evaluation

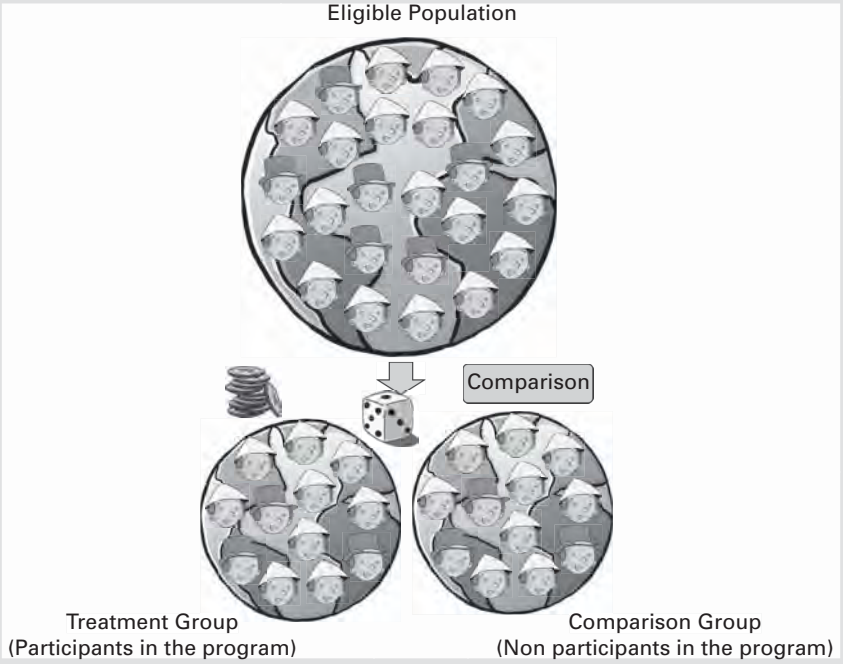
Confusion sometimes arises between random sampling and randomized assignment. What if someone proudly tells you that they are implementing an impact evaluation by interviewing a *random sample* of participants and nonparticipants? Assume that you observe a group of individuals participating in an employment program, and a group of individuals not participating in the program. What if you were to take a random sample of each of these two groups? The first figure illustrates that you would obtain a random sample of participants and a random sample of nonparticipants.

Figure B15.1.1 Random Sampling among Noncomparable Groups of Participants and Nonparticipants



(continued)

Figure B15.1.2 Randomized Assignment of Program Benefits between a Treatment Group and a Comparison Group



If participants and nonparticipants have different characteristics, so will the sample of participants and nonparticipants. Random sampling does not make two noncomparable groups comparable, and thus does not provide internal validity for the impact evaluation. This is why random sampling is not sufficient for impact evaluation.

As should be clear from the discussion in part 2, randomized assignment of program benefits is different from random sampling. The randomized assignment process starts from an eligible population of interest and uses a randomization procedure to assign units (usually consisting of people or groups of people, such as children in a school) from the eligible population to a treatment group that will receive an intervention and a comparison group that will not. The randomization process of a program in the second figure is different than the random sampling process described in the first figure. As discussed in part 2, when randomized assignment is well implemented, it contributes to the internal validity of the impact evaluation. Random sampling can be useful to ensure external validity, to the extent that the sample is randomly drawn from the population of interest.

In the rest of this chapter, we discuss how the size of the sample matters for the precision of the impact evaluation. As will become clear, relatively larger samples are needed to obtain precise estimates of the population characteristics. Larger samples are also needed to be able to obtain more precise estimates of differences between treatment groups and comparison groups, that is, to estimate the impact of a program.

Deciding on the Size of a Sample for Impact Evaluation: Power Calculations

As discussed, sampling describes the process of drawing a sample of units from a population of interest to estimate the characteristics of that population. Larger samples give more precise estimates of the population characteristics. Exactly how large do samples need to be for impact evaluation? The calculations to determine how large the sample must be are called power calculations. We discuss the basic intuition behind power calculations by focusing on the simplest case: an evaluation conducted using a randomized assignment method, testing the effectiveness of a program against a comparison group that does not receive an intervention, and assuming that noncompliance is not an issue.[2] We briefly discuss additional considerations beyond this simple case at the end of the chapter.

The Rationale for Power Calculations

Power calculations indicate the minimum sample size needed to conduct an impact evaluation and to convincingly answer the policy question of interest. In particular, power calculations can be used to

- Assess whether existing data sets are large enough to conduct an impact evaluation.
- Avoid collecting too little data. If the sample is too small, you may not be able to detect positive impact—even if it existed—and may thus conclude that the program had no effect. That could lead to a policy decision to eliminate the program, and that would be detrimental.
- Help make decisions about adequate sample size. Larger sample sizes provide more accurate estimates of program impacts, but collecting information can be very costly. Power calculations provide key inputs to assess trade-offs between costs required to collect additional data and gains from greater precision within the impact evaluation.

Key Concept

Power calculations provide an indication of the smallest sample with which it is possible to precisely estimate the impact of a program, that is, the smallest sample that will allow us to detect meaningful differences in outcomes between the treatment and comparison groups.

Power calculations provide an indication of the smallest sample (and lowest budget) with which it is possible to measure the impact of a program, that is, the smallest sample that will allow meaningful differences in outcomes between the treatment and comparison groups to be detected. Power calculations are thus crucial for determining which programs are successful and which are not.

As discussed in chapter 1, the basic evaluation question tackled by impact evaluations is, what is the impact or causal effect of a program on an outcome of interest? The simple hypothesis embedded in that question can be restated as follows: Is the program impact different from zero? In the case of randomized assignment, answering this question requires two steps:

1. Estimate the average outcomes for the treatment and comparison groups.
2. Assess whether a difference exists between the average outcome for the treatment group and the average outcome for the comparison group.

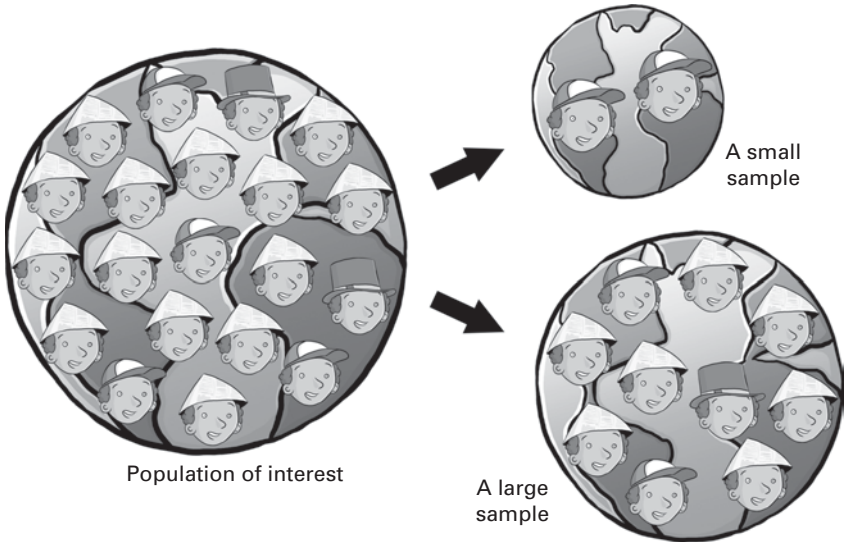
We now discuss how to estimate average outcomes for each group, and then how to test for a difference between groups.

Estimating Average Outcomes for the Treatment and Comparison Groups

Assume that you are interested in estimating the impact of a nutrition program on the weight of children at age two, and that 200,000 children are eligible for the program. From all eligible children, 100,000 were randomly assigned to participate in the program. The 100,000 eligible children who were not randomly assigned to the program serve as the comparison group. As a first step, you will need to estimate the average weight of the children who participated and the average weight of those who did not.

To determine the average weight of participating children, one could weigh every one of the 100,000 participating children and then average the weights. Of course, doing that would be extremely costly. Luckily, it is not necessary to measure every child. The average can be estimated using the average weight of a sample drawn from the population of participating children.[3] The more children in the sample, the closer the sample average will be to the true average. When a sample is small, the average weight constitutes a very imprecise estimate of the average in the population. For example, a sample of two children will not give a precise estimate. In contrast, a sample of 10,000 children will produce a more precise estimate that is much closer to the true average weight. In general, the more observations in the sample, the more precise the statistics obtained from the sample will be (figure 15.3).[4]

Figure 15.3 A Large Sample Is More Likely to Resemble the Population of Interest



So now we know that with a larger sample we provide a more precise and accurate image of the population of participating children. The same will be true for nonparticipating children: as the sample of nonparticipating children gets larger, we will know more precisely what that population looks like. But why should we care? If we are able to estimate the average outcome (weight) of participating and nonparticipating children more precisely, we will also be able to tell more precisely the difference in weight between the two groups—and that is the estimate of the impact of the program. To put it another way, if you have only a vague idea of the average weight of children in the participating (treatment) and nonparticipating (comparison) groups, then how can you have a precise idea of the difference in the weight of the two groups? That's right; you can't. In the following section, we will explore this idea in a slightly more formal way.

Comparing the Average Outcomes between the Treatment and Comparison Groups

Once you have estimated the average outcome (weight) for the treatment group (participating children selected by randomized assignment) and the comparison group (nonparticipating children selected by randomized assignment), you can proceed to determine whether the two outcomes

are different. This part is clear: you subtract the averages and check what the difference is. In statistical terms, the impact evaluation tests the *null (or default) hypothesis* against the *alternative hypothesis*.

The null hypothesis is the hypothesis that the program does not have an impact. It is expressed as:

H_0 : Impact or difference between outcomes in treatment and comparison groups = 0.

H_a : Impact or difference between outcomes in treatment and comparison groups $\neq$ 0.

Imagine that in the nutrition program example, you start with a sample of two treated children and two comparison children. With such a small sample, your estimate of the average weight of treated and comparison children, and thus your estimate of the difference between the two groups, will not be very reliable. You can check this by drawing different samples of two treated and two comparison children. What you will find is that the estimated impact of the program bounces around a lot.

By contrast, let us say that you start with a sample of 1,000 treated children and 1,000 comparison children. As discussed, your estimates of the average weight of both groups will be much more precise. Therefore, your estimate of the difference between the two groups will also be more precise.

For example, say that you find that the average weight in the sample of treatment (participating) children is 12.2 kilograms (kg), and the average in the sample of comparison (nonparticipating) children is 12.0 kg. The difference between the two groups is 0.2 kg. If these numbers came from samples of two observations each, you would not be very confident that the impact of the program is truly positive because the entire 0.2 kg could be due to the lack of precision in your estimates. However, if these numbers come from samples of 1,000 observations each, you would be more confident that you are quite close to the true program impact, which in this case would be positive.

The key question then becomes, Exactly how large must the sample be to allow you to know that a positive estimated impact is due to true program impact, rather than to lack of precision in your estimates?

Two Potential Errors in Impact Evaluations

When testing whether a program has an impact, two types of error can be made. A *type I error* is made when an evaluation concludes that a program has had an impact, when in reality it had no impact. In the case of the hypothetical nutrition intervention, this would happen if you, as a member

of the evaluation team, were to conclude that the average weight of the children in the treated sample is higher than that of the children in the comparison sample, even though the average weight of the children in the two populations is in fact equal and observed differences were purely coincidental. In this case, the positive impact you saw came purely from the lack of precision of your estimates.

A *type II error* is the opposite kind of error. A type II error occurs when an evaluation concludes that the program has had no impact, when in fact it has had an impact. In the case of the nutrition intervention, this would happen if you were to conclude that the average weight of the children in the two samples is the same, even though the average weight of the children in the treatment population is in fact higher than that of the children in the comparison population. Again, the impact should have been positive, but because of lack of precision in your estimates, you concluded that the program had zero impact.

When testing the hypothesis that a program has had an impact, statisticians can limit the size of type I errors. The likelihood of a type I error can be set by a parameter called the *significance level*. The significance level is often fixed at 5 percent—meaning that you can be 95 percent confident in concluding that the program has had an impact. If you are very concerned about committing a type I error, you can conservatively set a lower significance level—for example, 1 percent, so that you are 99 percent confident in concluding that the program has had an impact.

However, type II errors are also worrying for policy makers. Many factors affect the likelihood of committing a type II error, but the sample size is crucial. If the average weight of 50,000 treated children is the same as the average weight of 50,000 comparison children, then you probably can confidently conclude that the program has had no impact. By contrast, if a sample of two treatment children weigh on average the same as a sample of two comparison children, it is harder to reach a reliable conclusion. Is the average weight similar because the intervention has had no impact or because the data are not sufficient to test the hypothesis in such a small sample? Drawing large samples makes it less likely that you will observe only children who weigh the same simply by (bad) luck. In large samples, the difference in mean between the treated sample and comparison sample provides a better estimate of the true difference in mean between all treated and all comparison units.

The *power* (or *statistical power*) of an impact evaluation is the probability that it will detect a difference between the treatment and comparison groups, when in fact one exists. An impact evaluation has a high power if there is a low risk of not detecting real program impacts: that is, of committing a type II error. The previous examples show that the size of the

Key Concept

A *type I error* occurs when an evaluation concludes that a program has had an impact, when in reality it had no impact.
A *type II error* occurs when an evaluation concludes that the program has had no impact, when in fact it has had an impact.

Key Concept

Power is the probability of detecting an impact, when in fact one exists. An impact evaluation has high power if there is a low risk of not detecting real program impacts: that is, of committing a type II error.

sample is a crucial determinant of the power of an impact evaluation. The following sections will further illustrate this point.

Why Power Calculations Matter for Policy

The purpose of power calculations is to determine how large a sample is required to avoid concluding that a program has had no impact, when it has in fact had one (a type II error). The power of a test is equal to 1 minus the probability of a type II error.

An impact evaluation has *high power* if a type II error is unlikely to happen—meaning that you are unlikely to be disappointed by results showing that the program being evaluated has had no impact, when in reality it did have an impact.

From a policy perspective, *underpowered impact evaluations* with a high probability of type II errors are not only unhelpful but can also be very costly. A high probability of type II error jeopardizes the potential for an impact evaluation to identify statistically significant results. Putting resources into underpowered impact evaluations is therefore a risky investment.

Underpowered impact evaluations can also have serious practical consequences. For example, in the hypothetical nutrition intervention previously mentioned, if you were to conclude that the program was not effective, even though it was, policy makers might close down a program that, in fact, benefits children. It is therefore crucial to minimize the probability of type II errors by using large enough samples in impact evaluations. That is why carrying out power calculations is so crucial and relevant.

Power Calculations Step by Step

We now turn to the basic principles of power calculations, focusing on the simple case of a randomly assigned program. Carrying out power calculations requires examining the following five main questions:

1. Does the program operate through *clusters*?
2. What is/are the *outcome indicator(s)*?
3. What is the *minimum level of impact* that would justify the investment that has been made in the intervention?
4. What is the *mean of the outcome* for the population of interest? What is the *underlying variance* of the outcome indicator?
5. What are reasonable levels of *statistical power* and *statistical significance* for the evaluation being conducted?

Table 15.1 Examples of Clusters

Benefit	Level at which benefits are assigned (cluster)	Unit at which outcome is measured
Cash transfers	Village	Households
Malaria treatment	School	Individuals
Training program	Neighborhood	Individuals

Each of these questions applies to the specific policy context in which you have decided to conduct an impact evaluation.

The first step in power calculations is to determine whether the program that you want to evaluate creates any *clusters* through its implementation. An intervention whose level of intervention (often places) is different from the level at which you would like to measure outcomes (often people) creates clusters around the location of the intervention. For example, it may be necessary to implement a program at the hospital, school, or village level (in other words, through clusters), but you measure its impact on patients, students, or villagers (see table 15.1).[5] When an impact evaluation involves clusters, it is the number of clusters that largely determines the useful sample size. By contrast, the number of individuals within clusters matters less. We discuss this further below.

The nature of any sample data built from programs that are clustered is a bit different from that of samples obtained from programs that are not. As a result, power calculations will involve slightly different steps, depending on whether a program randomly assigns benefits among clusters or simply assigns benefits randomly among all units in a population. We will discuss each situation in turn. We start with the principles of power calculations when there are no clusters: that is, when the treatment is assigned at the level at which outcomes are observed. We then go on to discuss power calculations when clusters are present.

Power Calculations without Clusters

Assume that you have solved the first question by establishing that the program's benefits are not assigned by clusters. In other words, the program to be evaluated randomly assigns benefits among all units in an eligible population.

In the second step, you must identify the most important *outcome indicators* that the program was designed to improve. These indicators derive from the program objective, theory of change, and the fundamental evaluation research question, as discussed in part 1. Power calculations will also yield insights into the type of indicators for which impact evaluations can

identify impacts. Indeed, as we will further discuss, samples of varying sizes may be required to measure impacts on different indicators.

Third, you must determine the minimum impact that would justify the investment that has been made in the intervention. This is fundamentally a policy question, rather than a technical one. Is a cash transfer program a worthwhile investment if it reduces poverty by 5 percent, 10 percent, or 15 percent? Is an active labor market program worth implementing if it increases earnings by 5 percent, 10 percent, or 15 percent? The answer is highly specific to the context, but in all contexts it is necessary to determine the change in the outcome indicators that would justify the investment made in the program. Put another way, what is the level of impact below which an intervention should be considered unsuccessful? The answer to that question provides you with the *minimum detectable effect* that the impact evaluation needs to be able to identify. Answering this question will depend not only on the cost of the program and the type of benefits that it provides, but also on the opportunity cost of not investing funds in an alternative intervention.

Key Concept

The *minimum detectable effect* (MDE) is the effect size that an impact evaluation is designed to estimate for a given level of significance and power. All else being equal, larger samples are needed for an impact evaluation to detect smaller differences between the treatment and comparison groups, or to detect differences in a more variable outcome.

While minimum detectable effects can be based on policy objectives, other approaches can be used to establish them. It can be useful to benchmark minimum detectable effects against results from studies on similar programs to shed light on the magnitude of impacts that can be expected. For instance, education interventions often measure gains in terms of standardized test scores. Existing studies show that an increase in 0.1 standard deviation is relatively small, while an increase of 0.5 standard deviation is relatively large. Alternatively, ex ante simulations can be performed to assess the range of impacts that are realistic under various hypotheses. Examples of ex ante simulations were provided in chapter 1 for conditional cash transfer programs. Lastly, ex ante economic analysis can shed light on the size of the impacts that would be needed for the rate of return on a given investment to be sufficiently high. For instance, the annualized earnings gains triggered by a job training program would need to be above a prevailing market interest rate.

Intuitively, it is easier to identify a large difference between two groups than it is to identify a small difference between two groups. For an impact evaluation to identify a small difference between the treatment and comparison groups, a very precise estimate of the difference in mean outcomes between the two groups will be needed. This requires a large sample. Alternatively, for interventions that are judged to be worthwhile only if they lead to large changes in outcome indicators, the samples needed to conduct an impact evaluation will be smaller. Nevertheless, the minimum detectable effect should be set conservatively, since any impact smaller than the minimum desired effect is less likely to be detected.

Fourth, to conduct power calculations, you must ask an expert to estimate some basic parameters, such as a baseline mean and variance of the

outcome indicators. These benchmark values should preferably be obtained from existing data collected in a setting similar to the one where the program under study will be implemented, or from a pilot survey in the population of interest.[6] It is very important to note that the more variable the outcomes of interest prove to be, the larger the sample that will be needed to estimate a precise treatment effect. In the example of the hypothetical nutrition intervention, children's weight is the outcome of interest. If all individuals weigh the same at the baseline, it will be feasible to estimate the impact of the nutrition intervention in a small sample. By contrast, if baseline weights among children are very variable, then a larger sample will be required to estimate the program's impact.

Fifth, the evaluation team needs to determine a reasonable *power level* and *significance level* for the planned impact evaluation. As stated earlier, the power of a test is equal to 1 minus the probability of any type II error. Therefore, the power ranges from 0 to 1, with a high value indicating less risk of failing to identify an existing impact. A power of 0.8 is a widely used benchmark for power calculations. It means that you will find an impact in 80 percent of the cases where one has occurred. A higher level of power of 0.9 (or 90 percent) often provides a useful benchmark but is more conservative, increasing the required sample sizes.

The significance level is the probability of committing a type I error. It is usually set at 5 percent, so that you can be 95 percent confident in concluding that the program has had an impact if you do find a significant impact. Other common significance levels are 1 percent and 10 percent. The smaller your significance level, the more confident you can be that the estimated impact is real.

Once these five questions have been addressed, the power calculations expert can calculate the required sample size using standard statistical software.[7] The power calculation will indicate the required sample size, depending on the parameters established in steps 1 to 5. The computations themselves are straightforward, once policy-relevant parameters have been determined (particularly in steps 2 and 3).[8] If you are interested in the implementation of power calculations, the technical companion available on the book website provides examples of power calculations using Stata and Optimal Design.

When seeking advice from statistical experts, the evaluation team should ask for an analysis of the sensitivity of the power calculation to changes in the assumptions. That is, it is important to understand how much the required sample size will have to increase under more conservative assumptions (such as lower expected impact, higher variance in the outcome indicator, or a higher level of power). It is also good practice to commission power calculations for various outcome indicators, as the required sample

sizes can vary substantially if some outcome indicators are much more variable than others. Finally, the power calculations can also indicate the sample size needed to make comparison of program impacts across specific subgroups (for example, men or women, or other subgroups of the population of interest). Each subgroup would need to have the required sample size.

Evaluating the Impact of HISP: Deciding How Big a Sample Is Needed to Evaluate an Expanded HISP

Returning to our example in part 2, let us say that the ministry of health was pleased with the quality and results of the evaluation of the Health Insurance Subsidy Program (HISP). However, before scaling up the program, the ministry decides to pilot an expanded version of the program, which they call HISP+. The original HISP pays for part of the cost of health insurance for poor rural households, covering costs of primary care and drugs, but it does not cover hospitalization. The minister of health wonders whether an expanded HISP+ that also covers hospitalization would further lower out-of-pocket health expenditures of poor households. The ministry asks you to design an impact evaluation to assess whether HISP+ would decrease health expenditures for poor rural households.

In this case, choosing an impact evaluation design is not a challenge for you: HISP+ has limited resources and cannot be implemented universally immediately. As a result, you have concluded that randomized assignment would be the most viable and robust impact evaluation method. The minister of health understands how well the randomized assignment method can work and is supportive.

To finalize the design of the impact evaluation, you have hired a statistician who will help you establish how big a sample is needed. Before he starts working, the statistician asks you for some key inputs. He uses a checklist of five questions.

1. Will the HISP+ program generate clusters? At this point, you are not totally sure. You believe that it might be possible to randomize the expanded benefit package at the household level among all poor rural households that already benefit from HISP. However, you are aware that the minister of health may prefer to assign the expanded program at the village level, and that would create clusters. The statistician suggests conducting power calculations for a benchmark case without clusters, and then considering how results would change with clusters.

2. What is the outcome indicator? You explain that the government is interested in a well-defined indicator: out-of-pocket health expenditures of poor households. The statistician looks for the most up-to-date source to obtain benchmark values for this indicator and suggests using the follow-up survey from the HISP evaluation. He notes that among households that received HISP, the per capita yearly out-of-pocket health expenditures have averaged US$7.84.

3. What is the minimum level of impact that would justify the investment in the intervention? In other words, what decrease in out-of-pocket health expenditures below the average of US$7.84 would make this intervention worthwhile? The statistician stresses that this is not only a technical consideration, but truly a policy question; that is why a policy maker like you must set the minimum effect that the evaluation should be able to detect. You remember that based on ex ante economic analysis, the HISP+ program would be considered effective if it reduced household out-of-pocket health expenditures by US$2. Still, you know that for the purpose of the evaluation, it may be better to be conservative in determining the minimum detectable impact, since any smaller impact is unlikely to be captured. To understand how the required sample size varies based on the minimum detectable effect, you suggest that the statistician perform calculations for a minimum reduction of out-of-pocket health expenditures of US$1, US$2, and US$3.

4. What is the variance of the outcome indicator in the population of interest? The statistician goes back to the data set of treated HISP households, pointing out that the standard deviation of out-of-pocket health expenditures is US$8.

5. What would be a reasonable level of power for the evaluation being conducted? The statistician adds that power calculations are usually conducted for a power between 0.8 and 0.9. He recommends 0.9, but offers to perform robustness checks later for a less conservative level of 0.8.

Equipped with all this information, the statistician undertakes the power calculations. As agreed, he starts with the more conservative case of a power of 0.9. He produces the results shown in table 15.2.

The statistician concludes that to detect a US$2 decrease in out-of-pocket health expenditures with a power of 0.9, the sample needs to contain at least 672 units (336 treated units and 336 comparison units, with no clustering). He notes that if you were satisfied to detect a US$3 decrease in out-of-pocket health expenditures, a smaller sample of at least 300 units

Table 15.2 Evaluating HISP+: Sample Size Required to Detect Various Minimum Detectable Effects, Power = 0.9

Minimum detectable effect	Treatment group	Comparison group	Total sample
US$1	1,344	1,344	2,688
US$2	336	336	672
US$3	150	150	300

Note: The minimum detectable effect describes the minimum reduction of household out-of-pocket health expenditures that can be detected by the impact evaluation. Power = 0.9, no clustering.

Table 15.3 Evaluating HISP+: Sample Size Required to Detect Various Minimum Detectable Effects, Power = 0.8

Minimum detectable effect	Treatment group	Comparison group	Total sample
US$1	1,004	1,004	2,008
US$2	251	251	502
US$3	112	112	224

Note: The minimum detectable effect describes the minimum reduction of household out-of-pocket health expenditures that can be detected by the impact evaluation. Power = 0.8, no clustering.

(150 units in each group) would be sufficient. By contrast, a much larger sample of at least 2,688 units (1,344 in each group) would be needed to detect a US$1 decrease in out-of-pocket health expenditures.

The statistician then produces another table for a power level of 0.8. Table 15.3 shows that the required sample sizes are smaller for a power of 0.8 than for a power of 0.9. To detect a US$2 reduction in household out-of-pocket health expenditures, a total sample of at least 502 units would be sufficient. To detect a US$3 reduction, at least 224 units are needed. However, to detect a US$1 reduction, at least 2,008 units would be needed in the sample. The statistician stresses that the following results are typical of power calculations:

- The higher (more conservative) the level of power, the larger the required sample size.
- The smaller the impact to be detected, the larger the required sample size.

Table 15.4 Evaluating HISP+: Sample Size Required to Detect Various Minimum Desired Effects (Increase in Hospitalization Rate)
Power = 0.8, no clustering

Minimum detectable effect (percentage point)	Treatment group	Comparison group	Total sample
1	7,257	7,257	14,514
2	1,815	1,815	3,630
3	807	807	1,614

Note: The minimum desired effect describes the minimum change in the hospital utilization rate (expressed in percentage points) that can be detected by the impact evaluation.

The statistician asks whether you would like to conduct power calculations for other outcomes of interest. You suggest also considering the sample size required to detect whether HISP+ affects the hospitalization rate. In the sample of treated HISP villages, a household member visits the hospital in a given year in 5 percent of households; this provides a benchmark rate. The statistician produces a new table, which shows that relatively large samples would be needed to detect changes in the hospitalization rate (table 15.4) of 1, 2, or 3 percentage points from the baseline rate of 5 percent.

Table 15.4 shows that sample size requirements are larger for this outcome (the hospitalization rate) than for out-of-pocket health expenditures. The statistician concludes that if you are interested in detecting impacts on both outcomes, you should use the larger sample sizes implied by the power calculations performed on the hospitalization rates. If sample sizes from the power calculations performed for out-of-pocket health expenditures are used, the statistician suggests letting the minister of health know that the evaluation will not have sufficient power to detect policy-relevant effects on hospitalization rates.

HISP Question 8

A. Which sample size would you recommend to estimate the impact of HISP+ on out-of-pocket health expenditures?
B. Would that sample size be sufficient to detect changes in the hospitalization rate?

Power Calculations with Clusters

The previous discussion introduced the principles of carrying out power calculations for programs that do not create clusters. However, as discussed in part 2, some programs assign benefits at the cluster level. We now briefly describe how the basic principles of power calculations need to be adapted for clustered samples.

Key Concept

The number of clusters matters much more for power calculations than does the number of individuals within the clusters. At least 30 to 50 clusters are often required in each of the treatment and comparison groups, though sample size requirements will vary on a case-by-case basis, and power calculations are needed to ensure adequate sample size.

In the presence of clustering, an important guiding principle is that the number of clusters typically matters much more than the number of individuals within the clusters. A sufficient number of clusters is required to test convincingly whether a program has had an impact by comparing outcomes in samples of treatment and comparison units. It is the number of clusters that largely determines the useful or effective sample size. If you randomly assign treatment among a small number of clusters, the treatment and comparison clusters are unlikely to be identical. Randomized assignment between two districts, two schools, or two hospitals will not guarantee that the two clusters are similar. By contrast, randomly assigning an intervention among 100 districts, 100 schools, or 100 hospitals is more likely to ensure that the treatment and comparison groups are similar. In short, a sufficient number of clusters is necessary to ensure that balance is achieved. Moreover, the number of clusters also matters for the precision of the estimated treatment effects. A sufficient number of clusters is required to test the hypothesis that a program has an impact with sufficient power. When implementing an impact evaluation based on randomized assignment, it is therefore very important to ensure that the number of clusters is large enough.

You can establish the number of clusters required for precise hypothesis testing by conducting power calculations. Carrying out power calculations for cluster samples requires asking the same five questions listed above plus an additional one: How variable is the outcome indicator within clusters?

At the extreme, all outcomes within a cluster are perfectly correlated. For instance, it may be that household income is not especially variable within villages but that significant inequalities in income occur between villages. In this case, if you consider adding an individual to your evaluation sample, adding an individual from a new village will provide much more additional power than adding an individual from a village that is already represented. Since outcomes are fully correlated within a cluster, adding a new individual from the existing cluster will not add any new information. Indeed, in this case, the second villager is likely to look very similar to the original villager already included. In general, higher *intra-cluster correlation* in outcomes (that is, higher correlation in outcomes or characteristics between

units that belong to the same cluster) increases the number of clusters required to achieve a given power level.

In clustered samples, power calculations highlight the trade-offs between adding clusters and adding observations within clusters. The relative increase in power from adding a unit to a new cluster is almost always larger than that from adding a unit to an existing cluster. Although the gain in power from adding a new cluster can be dramatic, adding clusters may also have operational implications and increase the cost of program implementation or data collection. Later in this chapter, we show how to conduct power calculations with clusters in the case of HISP+ and discuss some of the trade-offs involved.

In many cases, at least 40 to 50 clusters in each treatment and comparison group are required to obtain sufficient power and guarantee balance of baseline characteristics when using randomized assignment methods. However, the number may vary depending on the various parameters already discussed, as well as the intra-cluster correlation. In addition, as will be discussed further below, the number will likely increase when using methods other than randomized assignment (assuming all else is constant).

Evaluating the Impact of HISP: Deciding How Big a Sample Is Needed to Evaluate an Expanded HISP with Clusters

After your first discussion with the statistician about power calculations for HISP+, you decided to talk briefly to the minister of health about the implications of randomly assigning the expanded HISP+ benefits among all individuals in the population who receive the basic HISP plan. The consultation revealed that such a procedure would not be politically feasible: in that context, it would be hard to explain why one person would receive the expanded benefits, while her neighbor would not.

Instead of randomization at the individual level, you therefore suggest randomly selecting a number of HISP villages to pilot HISP+. All villagers in the selected village would then become eligible. This procedure will create clusters and thus require new power calculations. You now want to determine how large a sample is required to evaluate the impact of HISP+ when it is randomly assigned by cluster.

You consult with your statistician again. He reassures you: only a little more work is needed. On his checklist, only one question is left

Table 15.5 Evaluating HISP+: Sample Size Required to Detect Various Minimum Detectable Effects (Decrease in Household Health Expenditures)
Power = 0.8, maximum of 100 clusters

Minimum detectable effect	Number of clusters	Units per cluster	Total sample with clusters	Total sample without clusters
US$1	100	102	10,200	2,008
US$2	90	7	630	502
US$3	82	3	246	224

Note: The minimum detectable effect describes the minimum reduction of household out-of-pocket health expenditures that can be detected by the impact evaluation. The number of clusters is the total number of clusters, half of which will be the number of clusters in the comparison group, and the other half the number of clusters in the treatment group.

unanswered. He needs to know how variable the outcome indicator is within clusters. Luckily, this is also a question he can answer using the HISP data. He finds that the within-village correlation of out-of-pocket health expenditures is equal to 0.04.

He also asks whether an upper limit has been placed on the number of villages in which it would be feasible to implement the new pilot. Since the program now has 100 HISP villages, you explain that you could have, at most, 50 treatment villages and 50 comparison villages for HISP+. With that information, the statistician produces the power calculations shown in table 15.5 for a power of 0.8.

The statistician concludes that to detect a US$2 decrease in out-of-pocket health expenditures, the sample must include at least 630 units: that is, 7 units per cluster in 90 clusters (45 clusters in the treatment group and 45 clusters in the comparison group). He notes that this number is higher than in the sample under randomized assignment at the household level, which required only a total of 502 units (251 in the treatment group and 251 in the comparison group; see table 15.3). To detect a US$3 decrease in out-of-pocket health expenditures, the sample would need to include at least 246 units, or 3 units in each of 82 clusters (41 clusters in the treatment group and 41 clusters in the comparison group).

The statistician then shows you how the total number of observations required in the sample varies with the total number of clusters. He decides to repeat the calculations for a minimum detectable effect of US$2 and a power of 0.8. The size of the total sample required to estimate such an effect increases strongly when the number of clusters diminishes

(table 15.6). With 120 clusters, a sample of 600 observations would be needed. If only 30 clusters were available, the total sample would need to contain 1,500 observations. By contrast, if 90 clusters were available, only 630 observations would be needed.

Table 15.6 Evaluating HISP+: Sample Size Required to Detect a US$2 Minimum Impact for Various Numbers of Clusters
Power = 0.8

Minimum detectable effect	Number of clusters	Units per cluster	Total sample with clusters
US$2	30	50	1,500
US$2	58	13	754
US$2	81	8	648
US$2	90	7	630
US$2	120	5	600

Note: The number of clusters is the total number of clusters, half of which will be the number of clusters in the comparison group and the other half the number of clusters in the treatment group. If the design did not have any clusters, 251 units in each group would be needed to identify a minimum detectable effect of US$2 (see table 15.3).

HISP Question 9

A. Which total sample size would you recommend to estimate the impact of HISP+ on out-of-pocket health expenditures?
B. In how many villages would you advise the minister of health to roll out HISP+?

Moving Beyond the Benchmark Case

In this chapter, we have focused on the benchmark case of an impact evaluation implemented using the randomized assignment method with full compliance. This is the simplest scenario, and therefore the most suitable to convey the intuition behind power calculations. Still, many practical aspects of power calculations have not been discussed, and deviations from the basic cases discussed here need to be considered carefully. Some of these deviations are discussed below.

Using quasi-experimental methods. All else being equal, quasi-experimental impact evaluation methods such as regression discontinuity, matching, or difference-in-differences tend to require larger samples than the randomized assignment benchmark. For instance, when using regression discontinuity

design, chapter 6 highlighted that only observations around the eligibility threshold can be used. A sufficiently large sample is required around that threshold. Power calculations are needed to estimate the required sample to make meaningful comparisons around the threshold.

On the other hand, the availability of several rounds of data can help increase the power of an impact evaluation for a given sample size. For instance, baseline data on outcomes and other characteristics can help make the estimation of the treatment effects more precise. The availability of repeated measures of outcomes after the start of the treatment can also help.

Examining different program modalities or design innovations. In the examples presented in this chapter, the total sample size was divided equally between treatment and comparison groups. In some cases, the main policy question of the evaluation may entail comparing program impacts between program modalities or design innovations. If this is the case, the expected impact may be relatively smaller than if a treatment group receiving a program were to be compared with a comparison group receiving no benefits at all. As such, the minimum desired effect between two treatment groups may be smaller than the minimum desired effect between a treatment and comparison group. The optimal distribution of the sample may lead to treatment groups that are relatively larger than the comparison group.[9] In impact evaluations with multiple treatment arms, power calculations may need to be implemented to separately estimate the size of each treatment and comparison group, depending on the main policy questions of interest.

Comparing subgroups. In other cases, some of the impact evaluation questions may focus on assessing whether program impacts vary between different subgroups, such as gender, age, or income categories. If this is the case, then sample size requirements will be larger, and power calculations will need to be adjusted accordingly. For instance, it may be that a key policy question is whether an education program has a larger impact on female students than on male students. Intuitively, you will need a sufficient number of students of each gender in the treatment group and in the comparison group to detect an impact for each subgroup. Setting out to compare program impacts between two subgroups can double the required sample size. Considering heterogeneity between more groups (for example, by age) can also substantially increase the size of the sample required. If such comparisons across groups are to be made in the context of an impact evaluation relying on randomized assignment, it is preferable to also take this into account when implementing the randomization, and in particular to perform randomization within blocks or strata (that is, within each of the subgroups to be compared). In practice, even if no comparison across subgroups is to be made, stratified or block randomization can help further maximize power for a given sample size.

Analyzing multiple outcomes. Particular care is needed when undertaking power calculations in cases where an impact evaluation will seek to test whether a program leads to changes in multiple outcomes. If many different outcomes are considered, there will be a relatively higher probability that the impact evaluation will find impacts on one of the outcomes just by chance. To address this, the impact evaluation team will need to consider testing for the joint statistical significance of changes in various outcomes. Alternatively, some indexes for families of outcomes can be constructed. These approaches to tackle multiple hypothesis testing have implications for power calculations and sample size, and as such need to be considered when determining the sample needed for the impact evaluation.[10]

Dealing with imperfect compliance or attrition. Power calculations often provide the minimum required sample size. In practice, implementation issues often imply that the actual sample size is smaller than the planned sample size. For instance, imperfect compliance may imply that only a share of the beneficiaries offered the program take it up. Sample size requirements increase when imperfect compliance arises. In addition, even if all individuals take up the program, some attrition may occur at the follow-up survey if not all individuals can be tracked. Even if such noncompliance or attrition is random and does not affect the consistency of the impact estimates, these aspects would affect the power of the impact evaluation. It is generally advisable to add a margin to the sample size predicted by power calculations to account for such factors. Similarly, data of lower quality will have more measurement error and make the outcomes of interest more variable, also requiring larger sample sizes.

The more advanced considerations mentioned in this section are beyond the scope of this book, but the additional resources listed at the end of this chapter can help. In practice, evaluation teams need to include or contract an expert who can perform power calculations, and the expert should be able to provide advice on more advanced issues.

Additional Resources

- For accompanying material to this chapter and hyperlinks to additional resources, please see the Impact Evaluation in Practice website (http://www.worldbank.org/ieinpractice).
- For examples of how to undertake power calculations in the Stata and Optimal Design software programs for the specific HISP case that illustrates this chapter, see the online technical companion available on the book website (http://www.worldbank.org/ieinpractice). The technical companion includes

additional technical material for readers with a background in statistics and econometrics.

- For detailed discussions of sampling (including other methods such as systematic sampling or multistage sampling) beyond the basic concepts discussed here, see the following resources:
  - Cochran, William G. 1977. *Sampling Techniques*, third edition. New York: John Wiley.
  - Kish, Leslie. 1995. *Survey Sampling.* New York: John Wiley.
  - Lohr, Sharon. 1999. *Sampling: Design and Analysis.* Pacific Grove, CA: Brooks Cole.
  - Thompson, Steven K. 2002. *Sampling*, second edition. New York: John Wiley.
  - Or, at a more basic level, Kalton, Graham. 1983. *Introduction to Survey Sampling.* Beverly Hills, CA: Sage.
- Practical guidance for sampling can be found in the following:
  - Grosh, Margaret, and Juan Muñoz. 1996. "A Manual for Planning and Implementing the Living Standards Measurement Study Survey." LSMS Working Paper 126, World Bank, Washington, DC.
  - UN (United Nations). 2005. *Household Sample Surveys in Developing and Transition Countries.* New York: United Nations.
  - Iarossi, Giuseppe. 2006. *The Power of Survey Design: A User's Guide for Managing Surveys, Interpreting Results, and Influencing Respondents.* Washington, DC: World Bank.
  - Fink, Arlene G. 2008. *How to Conduct Surveys: A Step by Step Guide*, fourth edition. Beverly Hills, CA: Sage.
- For a power calculation spreadsheet that will calculate the power for a given sample size after certain characteristics are inputted, see the Inter-American Development Bank Evaluation hub, in the Design section under Tools (www.iadb.org/evaluationhub).
- For more on power calculations and sample size, see the World Bank Impact Evaluation Toolkit, Module 3 on Design (Vermeersch, Rothenbühler, and Sturdy 2012). This module also includes a guide for doing ex ante power calculations, a paper about power calculations with binary variables, and a collection of useful references for further information on power calculations. (http://www.worldbank.org/health/impactevaluationtoolkit)
- For several blog posts about power calculations, see the World Bank Development Impact blog (http://blogs.worldbank.org/impactevaluations/).
- For a discussion of some considerations for power calculations in designs more complex than the benchmark case of randomized assignment in presence of perfect compliance, see the following:
  - Spybrook, Jessaca, Stephen Raudenbush, Xiaofeng Liu, Richard Congdon, and Andrés Martinez. 2008. *Optimal Design for Longitudinal and Multilevel Research: Documentation for the "Optimal Design" Software*. New York: William T. Grant Foundation.
  - Rosenbaum, Paul. 2009. "The Power of Sensitivity Analysis and Its Limit." Chapter 14 in *Design of Observational Studies,* by Paul Rosenbaum. New York: Springer Series in Statistics.

- On the topic of multiple hypothesis testing, see the following:
  - Duflo, E., R. Glennerster, M. Kremer, T. P. Schultz, and A. S. John. 2007. "Using Randomization in Development Economics Research: A Toolkit." Chapter 61 in *Handbook of Development Economics,* Vol. 4, 3895–962. Amsterdam: Elsevier.
  - Schochet, P. Z. 2008. *Guidelines for Multiple Testing in Impact Evaluations of Educational Interventions*. Prepared by Mathematica Policy Research Inc., for the Institute of Education Sciences, U.S. Department of Education, Washington, DC.
- A number of tools are available for those interested in exploring sample design further. For example, the W. T. Grant Foundation developed the freely available Optimal Design Software for Multi-Level and Longitudinal Research, which is useful for statistical power analysis in the presence of clusters. The Optimal Design software and manual can be downloaded at http://hlmsoft.net/od.

Notes

1. Strictly speaking, samples are drawn from sampling frames. In our discussion, we assume that the sampling frame perfectly overlaps with the population.
2. As discussed in part 2, compliance assumes that all the units assigned to the treatment group are treated and all the units assigned to the comparison group are not treated.
3. In this context, the term *population* does not refer to the population of the country, but rather to the entire group of children that we are interested in: the population of interest.
4. This intuition is formalized by a theorem called the *central limit theorem*. Formally, for an outcome y, the central limit theorem states that the sample mean $\bar{y}$ on average constitutes a valid estimate of the population mean. In addition, for a sample of size n and for a population variance σ^2, the variance of the sample mean is inversely proportional to the size of the sample:

$$var(\bar{y}) = \frac{\sigma^2}{n}$$

   As the size of the sample n increases, the variance of sample estimates tends to 0. In other words, the mean is more precisely estimated in large samples than in small samples.
5. The allocation of benefits by cluster is often made necessary by social or political considerations that make randomization within clusters impossible. In the context of an impact evaluation, clustering often becomes necessary because of likely spillovers, or contagion of program benefits between individuals within clusters. See discussion in chapter 11.
6. When computing power from a baseline, the correlation between outcomes over time should also be taken into account in power calculations.

7. For instance, Spybrook and others (2008) introduced Optimal Design, user-friendly software to conduct power calculations.
8. Having treatment and comparison groups of equal size is generally desirable. Indeed, for a given number of observations in a sample, power is maximized by assigning half the observations to the treatment group and half to the comparison group. However, treatment and comparison groups do not always have to be of equal size. See discussion at the end of the chapter.
9. The costs of the treatment can also be taken into consideration and lead to treatment and comparison group that are not of equal size. See, for instance, Duflo and others (2007).
10. See, for instance, Duflo and others (2007) or Schochet (2008).

References

Cochran, William G. 1977. *Sampling Techniques*, third edition. New York: John Wiley & Sons.

Duflo, E., R. Glennerster, and M. Kremer. 2007. "Using Randomization in Development Economics Research: A Toolkit." In *Handbook of Development Economics,* Vol. 4, edited by T. Paul Schultz and John Strauss, 3895–962. Amsterdam: Elsevier.

Fink, Arlene G. 2008. *How to Conduct Surveys: A Step by Step Guide*, fourth edition. Beverly Hills, CA: Sage.

Grosh, Margaret, and Paul Glewwe, eds. 2000. *Designing Household Survey Questionnaires for Developing Countries: Lessons from 15 Years of the Living Standards Measurement Study.* Washington, DC: World Bank.

Grosh, Margaret, and Juan Muñoz. 1996. "A Manual for Planning and Implementing the Living Standards Measurement Study Survey." LSMS Working Paper 126, World Bank, Washington, DC.

Iarossi, Giuseppe. 2006. *The Power of Survey Design: A User's Guide for Managing Surveys, Interpreting Results, and Influencing Respondents.* Washington, DC: World Bank.

Kalton, Graham. 1983. *Introduction to Survey Sampling.* Beverly Hills, CA: Sage.

Kish, Leslie. 1995. *Survey Sampling.* New York: John Wiley.

Lohr, Sharon. 1999. *Sampling: Design and Analysis.* Pacific Grove, CA: Brooks Cole.

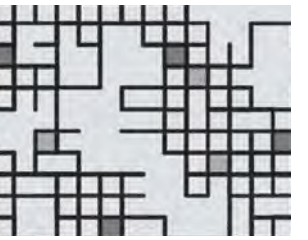
Rosenbaum, Paul. 2009. *Design of Observational Studies.* New York: Springer Series in Statistics.

Schochet, P. Z. 2008. *Guidelines for Multiple Testing in Impact Evaluations of Educational Interventions*. NCEE 2008-4018. National Center for Eduational Evaluation and Regional Assistance, Institute of Education Sciences. Washington, DC: U.S. Department of Education.

Spybrook, Jessaca, Stephen Raudenbush, Xiaofeng Liu, Richard Congdon, and Andrés Martinez. 2008. *Optimal Design for Longitudinal and Multilevel Research: Documentation for the "Optimal Design" Software*. New York: William T. Grant Foundation.

Thompson, Steven K. 2002. *Sampling*, second edition. New York: John Wiley.

Vermeersch, Christel, Elisa Rothenbühler, and Jennifer Sturdy. 2012. *Impact Evaluation Toolkit: Measuring the Impact of Results-Based Financing on Maternal and Child Health*. World Bank, Washington, DC. http://www.worldbank.org/health/impactevaluationtoolkit.



CHAPTER 16

Finding Adequate Sources of Data

Kinds of Data That Are Needed

In this chapter, we discuss the various sources of data that impact evaluations can use. We first discuss sources of existing data, particularly administrative data, and provide some examples of impact evaluations that have leveraged existing data. Since many evaluations require the collection of new data, we then discuss the steps in collecting new survey data. A clear understanding of these steps will help ensure that your impact evaluation is based on quality data that do not compromise the evaluation design. As a first step, you will need to commission the development of an appropriate questionnaire. In parallel, you will need to get help from a firm or government agency that specializes in data collection. The data collection entity will recruit and train field staff and pilot test the questionnaire. After making the necessary adjustments, the firm or agency will be able to proceed with fieldwork, collect the data, and digitize and process them, before they can be delivered, stored, and analyzed by the evaluation team.

Good quality data are required to assess the impact of the intervention on the outcomes of interest. The results chain discussed in chapter 2 provides a basis to define which indicators should be measured and when. Indicators are needed across the results chain.

Key Concept

Indicators are needed across the results chain to measure final outcomes, intermediate outcomes, as well as program benefits and quality of implementation.

Data about outcomes. The first and foremost need is data on outcome indicators directly affected by the program. Outcome indicators relate to the objectives the program seeks to achieve. As discussed in chapter 2, outcome indicators should preferably be selected so that they are SMART: specific, measurable, attributable, realistic, and targeted. The impact evaluation should not measure only those outcomes for which the program is directly accountable, however. Data on outcome indicators that the program indirectly affects, or indicators that capture unintended program effects, will maximize the value of the information that the impact evaluation generates, as well as the understanding of the program's overall effectiveness.

Data about intermediate outcomes. In addition, data on intermediary outcomes are useful to help understand the channels through which the program evaluated has impacted—or has not impacted—the final outcomes of interest. Impact evaluations are typically conducted over several time periods, and you must determine when to measure the outcome indicators. Following the results chain, you can establish a hierarchy of outcome indicators, ranging from short-term indicators that can be measured while participants are still in the program, such as school attendance collected in a short-term follow-up survey in the context of an education program, to longer-term ones, such as student achievement or labor market outcomes that can be measured in a longer-term follow-up survey after participants have exited the program. To measure impact convincingly over time, data are needed starting at a baseline before the program or innovation being evaluated is implemented. The section in chapter 12 on the timing of evaluations sheds light on when to collect data.

As we discussed in chapter 15 in the context of power calculations, some indicators may not be amenable to impact evaluation in small samples. Detecting impacts for outcome indicators that are extremely variable, that are rare events, or that are likely to be only marginally affected by an intervention may require prohibitively large samples. For instance, identifying the impact of an intervention on maternal mortality rates will be feasible only in a sample that contains tens of thousands of pregnant women, since mortality is (thankfully) a rare event. In such a case, it may be necessary to refocus the impact evaluation on more intermediate indicators, related to the final outcomes, but for which there is sufficient power to detect effects. In the case of an intervention meant to reduce maternal mortality, an intermediate indicator may be related to health service utilization during pregnancy and institutional delivery, which are associated with mortality. The power calculations

discussed in chapter 15 can help shed light on the indicators on which impacts can be detected, and those on which impacts may be harder to detect without very large samples.

Data about program activities and outputs. Indicators are also required for the part of the results chain that describes program activities and outputs. In particular, *program monitoring data* can provide essential information about the delivery of the intervention, including who the beneficiaries are and which program benefits or outputs they may have received. At minimum, monitoring data are needed to know when a program starts and who receives benefits, as well as to provide a measure of the intensity or quality of the intervention. This is particularly important in cases when the program may not be delivered to all beneficiaries with the same content, quality, or duration. A good understanding of the extent to which the intervention has been delivered as designed is essential to interpret the impact evaluation results, including whether they highlight the effectiveness of the program implemented as designed or with some flaws in implementation.

Additional data. Other data required by the impact evaluation can depend on the methodology used. Data on other factors that may affect the outcome of interest may be needed to control for outside influences. This aspect is particularly important when using evaluation methods that rely on more assumptions than randomized methods do. Sometimes it is also necessary to have data on outcomes and other factors over time to calculate trends, as is the case with the difference-in-differences method. Accounting for other factors and past trends also helps increase statistical power. Even with randomized assignment, data on other characteristics can make it possible to estimate treatment effects more precisely. They can be used to include additional controls or analyze the heterogeneity of the program's effects along relevant characteristics.

The design selected for the impact evaluation will also affect the data requirements. For example, if either the matching or the difference-in-differences method is chosen, it will be necessary to collect data on a broad array of characteristics for both treatment and comparison groups, making it possible to carry out a range of robustness tests, as described in part 2 or chapter 11 (see table 11.2).

For each evaluation, it is useful to develop a matrix that lists the question of interest, the outcome indicators for each question, the other types of indicators needed, and the source of data, as outlined in figure 2.1 on the results chain in chapter 2. The preparation of an impact evaluation plan and pre-analysis plan are other important opportunities to define a precise list of key indicators required for the impact evaluation.

Using Existing Quantitative Data

One of the first questions to consider when designing the impact evaluation is what sources of data it will use. A fundamental consideration is whether the impact evaluation will rely on existing data or require the collection of new data.

Some existing data are almost always needed at the outset of an impact evaluation to estimate benchmark values of indicators or to conduct power calculations, as discussed in chapter 15. Beyond the planning stages, the availability of existing data can substantially diminish the cost of conducting an impact evaluation. While existing data, and in particular administrative data, are probably underused in impact evaluation in general, the feasibility of using existing data for impact evaluation needs to be carefully assessed.

Indeed, as discussed in chapter 12, data collection is often the largest cost when implementing an impact evaluation. However, to determine whether existing data can be used in a given impact evaluation, a range of questions must be considered:

- *Sampling*. Are existing data available for both the treatment and comparison groups? Are existing samples drawn from a sampling frame that coincides with the population of interest? Were units drawn from the sampling frame based on a probabilistic sampling procedure?
- *Sample size*. Are existing data sets large enough to detect changes in the outcome indicators with sufficient power? The answer to this question depends on the choice of the outcome indicators, as well as on the results of the power calculations discussed in chapter 15.
- *Availability of baseline data*. Are the existing data available for both the treatment and comparison groups prior to the rollout of the program or innovation to be evaluated? The availability of baseline data is important to document balance in preprogram characteristics between treatment and comparison groups when randomized methods are used, and critical for the implementation of quasi-experimental designs.
- *Frequency*. Are the existing data collected frequently enough? Are they available for all units in the sample over time, including for the times when the outcome indicators need to be measured according to the results chain and the logic of the intervention?
- *Scope*. Do existing data contain all the indicators needed to answer the policy questions of interest, including the main outcome indicators and the intermediary outcomes of interest?

- *Linkages to program monitoring information.* Can existing data be linked to monitoring data on program implementation, including to observe which units are in the treatment and comparison groups, and whether all units assigned to the treatment group received the same benefits?
- *Unique identifiers.* Do unique identifiers exist to link across data sources?

As the questions above highlight, the requirements for existing data are quite significant, and it is not common for existing data to be sufficient for impact evaluations. Still, with the rapid growth in the scope and coverage of information systems, as well as the overall evolution toward a world where digital data from a broad range of sources are routinely stored, an increasing number of impact evaluations can consider the use of existing data. A range of potential sources of existing data can be used for impact evaluation, including census data, national surveys, or administrative data.

Population census data can provide comprehensive data for the complete population. They can be used in impact evaluations when they are available at a sufficiently disaggregated level and include details to know which units are in a treatment or comparison group, such as geographic or personal identifiers. Census data are fielded infrequently and usually include only a small set of key indicators. However, census data are sometimes collected to feed into information systems or registries that provide the basis to target public programs, including unique identifiers that can support linkages with other existing data sets.

Nationally representative surveys such as household surveys, living standards measurement surveys, labor force surveys, demographic and health surveys, enterprise surveys, or facility surveys can also be considered. They may contain a comprehensive set of outcome variables, but they rarely contain enough observations from both the treatment and comparison groups to conduct an impact evaluation. Assume, for example, that you are interested in evaluating a large national program that reaches 10 percent of the households in a given country. If a nationally representative survey is carried out on 5,000 households every year, it may contain roughly 500 households that receive the program in question. Is this sample large enough to conduct an impact evaluation? Power calculations can answer this question, but in many cases the answer is no.

In addition to exploring whether you can use existing surveys, you should also find out if any new national data collection efforts are being planned. If a survey is planned that will cover the population of interest, you may also be able to introduce a question or series of questions as part of that survey. If a survey measuring the required indicators is already planned, there might be opportunities to oversample a particular population to ensure appropriate coverage of the treatment and comparison groups and

accommodate the impact evaluation. For instance, the evaluation of the Nicaraguan Social Fund complemented a national Living Standards Measurement Study with an extra sample of beneficiaries (Pradhan and Rawlings 2002).

Administrative data are typically collected by public or private agencies as part of their regular operations, usually relatively frequently, and often to monitor the services delivered or record interactions with users. In some cases, administrative data contain outcome indicators needed for impact evaluation. For instance, education systems gather school records on students' enrollment, attendance, or test scores, and can also compile information on school inputs and teachers. Similarly, health systems can collect data on the characteristics and location of health centers, the supply of health services, and the allocation of resources. They can also consolidate data collected in health centers on patients' medical records, anthropometrics, and vaccination histories, and more broadly, data on the incidence of diseases and vital statistics. Public utility agencies collect data on water or electricity use. Tax agencies may collect data on income and taxes. Transport systems collect data on passengers and travel times. Financial system agencies collect data on customers' transactions or credit history. All these sources of existing data can potentially be used for impact evaluations. They sometimes include long time series that can help track units over time.

An assessment of data availability and quality is critical when considering whether to rely on administrative data. In some cases, data from administrative sources may be more reliable than survey data. For instance, a study in Malawi found that respondents overstated school attendance and enrollment in self-reported data from a household survey compared with administrative records obtained in schools; thus impact evaluation results were more reliable if based on the administrative data (Baird and Özler 2012). At the same time, in many contexts, administrative data are collected by a large number of providers and can be of unequal quality. Thus their reliability needs to be fully assessed before a decision is made to rely on administrative data for the impact evaluation. One critical aspect is to ensure that single identifiers exist to connect administrative data with other data sources, including program monitoring data documenting which units have received program benefits. When such identifiers exist—such as national identification numbers used consistently—a large amount of work to prepare and clean data may be avoided. In all cases, the protection of confidentiality is an important part of the data preparation and data management protocol. The ethical principles to protect human subjects (see discussion in chapter 13) also apply to the use of existing data.

Some influential retrospective evaluations have relied on administrative records (Galiani, Gertler, and Schargrodsky [2005] on water policy in Argentina; Ferraz and Finan [2008] on audits and politicians' performance; and Chetty, Friedman, and Saez [2013] on tax credits in the United States). Box 16.1 provides an example of a health impact evaluation in Argentina. Box 16.2 illustrates the use of administrative data in the impact evaluation of a cash transfer program in Honduras.

In some cases, the data required for impact evaluation can be collected by rolling out new information or administrative data systems. Such roll-out can be coordinated with the implementation of an evaluation design, so that outcome indicators are collected for a treatment and a comparison group at multiple times. The setup of information systems may need to be established before new interventions are launched, so that

Box 16.1: Constructing a Data Set in the Evaluation of Argentina's Plan Nacer

When evaluating Argentina's results-based health-financing program, Plan Nacer, Gertler, Giovagnoli, and Martinez (2014) combined administrative data from several sources to form a large and comprehensive database for analysis. After several previous evaluation strategies were unsuccessful, the researchers turned to an instrumental variables approach. This required a substantial amount of data from the universe of all birth records in the seven provinces studied.

The researchers needed data on prenatal care and birth outcomes, which could be found in birth registries at public hospitals. Then they needed to determine whether the mother was a beneficiary of Plan Nacer and whether the clinic she visited was incorporated into the program at the time of the visit. To construct a database with all this information, the evaluation team linked five different data sources, including public maternity hospital databases, Plan Nacer program implementation data, pharmaceutical records, the 2001 population census, and geographic information for health facilities. Obtaining medical records on individual births at maternity hospitals was among the most challenging tasks. Each maternity hospital collected prenatal care and birth outcome data, but only about half the records were digitized. The rest were on paper, requiring the evaluation team to merge the paper records into the computerized system.

Overall, the team was able to compile a comprehensive database for 78 percent of births occurring during the evaluation period. This yielded a large data set that allowed them to examine the impact of Plan Nacer on relatively rare events, such as neonatal mortality. This is typically not possible in evaluations with smaller samples collected through surveys. The evaluation found that beneficiaries of Plan Nacer had a 74 percent lower chance of in-hospital neonatal mortality than nonbeneficiaries.

Source: Gertler, Giovagnoli, and Martinez 2014.

Box 16.2: Using Census Data to Reevaluate the PRAF in Honduras

Honduras's Programa de Asignación Familiar (PRAF) aimed at improving educational and health outcomes for young children living in poverty. It provided cash transfers to eligible households conditional on regular school attendance and health center visits. The program began in 1990. An evaluation component was included in the second phase of the PRAF in 1998. Glewwe and Olinto (2004) and Morris and others (2004) reported positive impacts on education and health outcomes.

Several years later, Galiani and McEwan (2013) reevaluated the impact of the program using a different source of data. While the original impact evaluation collected survey data from 70 out of 298 municipalities, Galiani and McEwan used data from the 2001 Honduran census. They merged individual and household-level data from the census with municipal-level data on the treatment communities. This provided the researchers with a larger sample size that allowed them to test the robustness of the findings, in addition to spillover effects. Moreover, since the researchers had census data from all the municipalities, they were able to apply two different regression discontinuity designs (RDDs) using alternate comparison groups. For the first RDD, the researchers used the eligibility cutoff; for the second, they used municipal borders.

Like the previous impact evaluations, Galiani and McEwan found positive and statistically significant impacts from the program. However, their estimates implied that the PRAF had a much larger impact than the impact found in the original evaluation. They found that the PRAF increased school enrollment for eligible children by 12 percent more than those in the comparison group. The results from the alternate regression discontinuity designs generally confirmed the robustness of the findings.

Source: Galiani and McEwan 2013.

administrative centers in the comparison group use the new information system before receiving the intervention to be evaluated. Because the quality of administrative data can vary, auditing and external verification are required to guarantee the reliability of the evaluation. Collecting impact evaluation data through administrative sources instead of through surveys can dramatically reduce the cost of an evaluation, but it may not always be feasible.

Even if existing data are not sufficient for an entire impact evaluation, they can sometimes be used for parts of the impact evaluation. For example, in some cases, programs collect detailed targeting data on potential beneficiaries to establish who is eligible. Or census data may be available shortly before a program is rolled out. In such cases, the existing data can sometimes be used to document baseline balance in preprogram

characteristics in the treatment and comparison groups, even though additional follow-up data would still need to be collected to measure a broader set of outcome indicators.

Collecting New Survey Data

Only in relatively rare cases are existing data sufficient for an entire impact evaluation. If administrative data are not sufficient for your evaluation, you will likely have to rely on survey data. As a result, you will most likely have to budget for the collection of new data. Although data collection is often the major cost for an impact evaluation, it can also be a high-return investment upon which the quality of the evaluation often depends. The collection of new data provides the flexibility to ensure that all the necessary indicators are measured for a comprehensive assessment of program performance.

Most impact evaluations require survey data to be collected, including at least a *baseline survey* before the intervention or innovation to be evaluated, and a *follow-up survey* after it has been implemented. Survey data may be of various types, depending on the program to be evaluated and the unit of analysis. For instance, enterprise surveys use firms as the main unit of observation, facility surveys use health centers or schools as the main unit of observation, and household surveys use households as the main unit of observation. Most evaluations rely on individual or household surveys as a primary data source. In this section, we review some general principles of collecting survey data. Even though they primarily relate to household surveys, the same principles also apply to most other types of survey data.

The first step in deciding whether to use existing data or collect new survey data will be to determine the sampling approach, as well as the size of the sample that is needed (as discussed in chapter 15). Once you decide to collect survey data for the evaluation, you will need to

- Determine who will collect the data,
- Develop and pilot the data collection instrument,
- Conduct fieldwork and undertake quality control, and
- Process and store the data.

The implementation of those various steps is usually commissioned, but understanding their scope and key components is essential to managing a quality impact evaluation effectively.

Determining Who Will Collect the Data

You will need to designate the agency in charge of collecting data early on. Some important trade-offs must be considered when you are deciding who should collect impact evaluation data. Potential candidates for the job include

- The institution in charge of implementing the program,
- Another government institution with experience collecting data (such as a national statistical agency), or
- An independent firm or think tank that specializes in data collection.

The data collection entity always needs to coordinate closely with the agency implementing the program. Close coordination is required to ensure that no program operations are implemented before baseline data have been collected. When baseline data are needed for the program's operation (for instance, data for an eligibility index, in the context of an evaluation based on a regression discontinuity design), the entity in charge of data collection must be able to process the data quickly and transfer the data to the institution in charge of program operations. Close coordination is also required in timing the collection of follow-up survey data. For instance, if you have chosen a randomized rollout, the follow-up survey must be implemented before the program is rolled out to the comparison group, to avoid contamination.

Key Concept

The same data collection procedures should be used for both the comparison and treatment groups.

An extremely important factor in deciding who should collect data is that the same data collection procedures should be used for both the comparison and treatment groups. The implementing agency often has contact only with the treatment group and so is not in a good position to collect data for the comparison groups. But using different data collection agencies for the treatment and comparison groups is very risky, as it can create differences in the outcomes measured in the two groups simply because the data collection procedures differ. If the implementing agency cannot collect data effectively for both the treatment and comparison groups, the possibility of engaging an external institution or agency to do so should be strongly considered.

In some contexts, it may also be advisable to commission an independent agency to collect data to ensure that the data are considered objective. Concerns that the program-implementing agency does not collect objective data may not be warranted, but an independent data collection body that has no stake in the evaluation results can add credibility to the overall impact evaluation effort. It may also ensure that respondents do not perceive the survey to be part of the program and thus may minimize the risk that respondents will give strategic responses in an attempt to increase their perceived chances to participate in a program.

Because data collection involves a complex sequence of operations, it is recommended that a specialized and experienced entity be responsible for it. Few program-implementing agencies have sufficient experience to collect the large-scale, high-quality data necessary for an impact evaluation. In most cases, you will have to consider commissioning a local institution, such as a national statistical agency or a specialized firm or think tank.

Commissioning a local institution such as a national statistical agency can give the institution exposure to impact evaluation studies and help build its capacity—which may in itself be a side benefit of the impact evaluation. However, national statistical agencies may not always have the logistical capacity to take on extra mandates in addition to their regular activities. They may also lack the necessary experience in fielding surveys for impact evaluations, such as experience in successfully tracking individuals over time or in implementing nontraditional survey instruments. If such constraints appear, contracting an independent firm or think tank specialized in data collection may be more practical.

You do not necessarily have to use the same entity to collect information at baseline and in follow-up surveys, which may vary in scope. For instance, for an impact evaluation of a training program, for which the population of interest comprises the individuals who signed up for the course, the institution in charge of the course could collect the baseline data when individuals enroll. It is unlikely, however, that the same agency will also be the best choice to collect follow-up information for both the treatment and comparison groups. In this context, contracting rounds of data collection separately has its advantages, but efforts should be made not to lose any information between rounds that will be useful in tracking households or individuals, as well as to ensure that baseline and follow-up data are measured consistently.

To determine the best institution for collecting impact evaluation data, all these factors—experience in data collection, ability to coordinate with the program's implementing agency, independence, opportunities for capacity building, adaptability to the impact evaluation context—must be weighed, together with the expected cost and likely quality of the data collected in each case. One effective way to identify the organization best placed to collect quality data is to write clear terms of reference and ask organizations to submit technical and financial proposals.

Because the prompt delivery and the quality of the data are often crucial for the reliability of the impact evaluation, the contract for the agency in charge of data collection must be structured carefully. The scope of the expected work and deliverables must be made extremely clear. In addition, it is often advisable to introduce incentives into contracts and link those incentives to clear indicators of data quality. For instance, the nonresponse

rate is a key indicator of data quality. To create incentives for data collection agencies to minimize nonresponse, the contract can stipulate one unit cost for the first 80 percent of the sample, a higher unit cost for the units between 80 percent and 90 percent, and again a higher unit cost for units between 90 percent and 100 percent. Alternatively, a separate contract can be written for the survey firm to track nonrespondents. In addition, the data collection contract may include incentives or conditions related to verification of data quality, such as through back-checks or quality audits among a subsample of the impact evaluation survey.

Developing and Piloting the Data Collection Instrument

When commissioning data collection, the evaluation team has an important role to play in providing specific guidance on the content of the data collection instruments or questionnaires. Data collection instruments must elicit all the information required to answer the policy question set out by the impact evaluation. As we have discussed, *indicators* must be measured throughout the results chain, including indicators for final outcomes, intermediate outcomes, and measures of program benefits and quality of implementation.

It is important to be selective about which indicators to measure. Being selective helps limit data collection costs, simplifies the task of the data collection agency, and improves the quality of the data collected by minimizing demands on the enumerators and the respondents' time. Collecting information that is either irrelevant or unlikely to be used has a very high cost. Additional data require more time for preparing, training, collecting, and processing. With limited availability and attention spans, respondents may provide decreasing quality information as the survey drags on, and interviewers will have added incentives to cut corners to meet their survey targets. Thus extraneous questions are not "free." Having clear objectives for the impact evaluation that are aligned with well-defined program objectives can help you prioritize necessary information. A preanalysis plan written in advance (see discussion in chapters 12 and 13) will help ensure that the survey collects the data required for the impact analysis and avoids the inclusion of extraneous (and costly) additional information.

It is preferable to collect data on outcome indicators and control characteristics consistently at baseline and at follow-up. Having baseline data is highly desirable. Even if you are using randomized assignment or a regression discontinuity design, where simple postintervention differences can in principle be used to estimate a program's impact, baseline data are essential for testing whether the design of the impact evaluation is adequate (see discussion in part 2). Having baseline data can give you an insurance

policy when randomization does not work, in which case the difference-in-differences method can be used instead. Baseline data are also useful during the impact analysis stage, since baseline control variables can help increase statistical power and allow you to analyze impacts on different subpopulations. Finally, baseline data can be used to enhance the design of the program. For instance, baseline data sometimes make it possible to analyze targeting efficiency or to provide additional information about beneficiaries to the agency implementing the program. In some cases, the follow-up survey may include a broader set of indicators than the baseline survey.

Once you have defined the core data that need to be collected, the next step is to determine exactly how to measure those indicators. *Measurement* is an art in itself and is best handled by specialists, including the impact evaluation research team, the agency hired to collect data, survey experts, and experts in the measurement of specific complex indicators. Outcome indicators should be as consistent as possible with local and international best practice. It is always useful to consider how indicators of interest have been measured in similar surveys both locally and internationally. Using the same indicators (including the same survey modules or questions) ensures comparability between the preexisting data and the data collected for the impact evaluation. Choosing an indicator that is not fully comparable or not well measured may limit the usefulness of the evaluation results. In some cases, it may make sense to invest the resources to collect the new innovative outcome indicator, as well as a more established alternative.

Key Concept

Measuring indicators is an art and is best handled by specialists, including the impact evaluation research team, the agency hired to collect data, survey experts, and experts in the measurement of specific complex indicators.

Particular attention should be paid to ensuring that all the indicators can be measured in exactly the same way for all units in both the treatment group and the comparison group. Using different data collection methods (for example, using a phone survey for one group and an in-person survey for the other) creates the risk of generating bias. The same is true of collecting data at different times for the two groups (for example, collecting data for the treatment group during the rainy season and for the comparison group during the dry season). That is why the procedures used to measure any outcome indicator should be formulated very precisely. The data collection process should be exactly the same for all units. Within a questionnaire, each module related to the program should be introduced without affecting the flow or framing of responses in other parts of the questionnaire. In fact, when possible, it is best to avoid making any distinction between treatment and comparison groups in the data collection process. In most cases, the agency conducting the data collection (or at least the individual surveyors) should not have a reason to know the treatment or comparison status of the individuals in the survey.

One important decision to make is how to measure the outcome indicators, including whether through traditional questionnaire-based surveys and self-reported questions or through other methods. In recent years, several advances have been made in measuring key outcomes or behaviors that are relevant for impact evaluation. Advances include refining methods to collect self-reported data through questionnaires, as well as techniques to measure key outcomes directly.

Questionnaire design has been the subject of significant research. Entire books have been written about how best to measure particular indicators in specific contexts, including on the way to phrase questions asked in household surveys.[1] There is also a growing evidence base on how best to design questionnaires to collect agricultural data, consumption data, or employment data to maximize their precision.[2] Some recent evidence comes from randomized experiments testing different ways of structuring questionnaires and comparing their reliability.[3] Accordingly, questionnaire design requires attention to international best practice, as well as local experiences in measuring indicators. Small changes in the wording or sequencing of questions can have substantial effects in the data collected, so that great attention to details is essential in questionnaire development. This is especially important when attempting to ensure comparability across surveys, including, for instance, to measure outcomes repeatedly over time. Box 16.3 discusses guidelines related to questionnaire design and provides additional references.

A growing set of techniques has been developed to obtain *direct measurement of outcomes*. For instance, in the health sector, vignettes are sometimes used to present particular symptoms to health workers and to assess whether the provider recommends the appropriate treatment based on established guidelines and protocols. Such vignettes provide a direct measure of health providers' knowledge. Recent evaluations are relying on standardized patients (also known as incognito or simulated patients) to visit health centers and directly assess the quality of services delivered.[4] In the education sector, many evaluations seek to assess program impacts on students' learning. To do so, a range of learning assessments or direct measures of students' skills is used. Similarly, various test batteries have been developed to directly measure cognitive, linguistic, or motor development among young children in the context of impact evaluations of early childhood development interventions. Progress has also been made to obtain direct measures of skills among adults, including socioemotional skills or personality traits. Besides direct measurement of skills, a growing number of impact evaluations seek to obtain measures of teaching quality through direct observations of teachers' behaviors in the classroom.

Box 16.3: Designing and Formatting Questionnaires

Although questionnaire design in impact evaluations is integral to the quality of the data, it is often overlooked. Designing a questionnaire is a complex, long, and iterative process involving many decisions along the way about what can be measured and how. The applied impact evaluation methods course at the University of California, Berkeley (http://aie.cega.org) provides a guide to questionnaire design, outlining three phases: content, drafting, and testing. Throughout these phases, the module highlights the importance of involving relevant stakeholders, allowing enough time for repeated iterations and careful testing:

1. *Content.* Determine the content of a survey by first defining the effects that need to be measured, the observation units, and correlations with other factors. These conceptual definitions will then need to be translated into concrete indicators.
2. *Drafting.* Draft questions to measure the selected indicators. This is a critical step, as the quality of the data relies on it. The module provides more in-depth recommendations on the wording of questions, the organization of the survey, formatting, and other key considerations.
3. *Testing.* Test the questionnaire on three levels: the question, the module, and the whole survey.

The format of the questionnaire is also important to ensure quality data. Because different ways of asking the same survey question can yield different answers, both the framing and the format of the questions should be the same for all units to prevent any respondent or enumerator bias. UN (2005) makes six specific recommendations regarding the formatting of questionnaires for household surveys. These recommendations apply equally to most other data collection instruments:

1. Each question should be written out in full in the questionnaire, so that the interviewer can conduct the interview by reading each question word for word.
2. The questionnaire should include precise definitions of all the key concepts used in the survey, so that the interviewer can refer to the definition during the interview if necessary.
3. Each question should be as short and simple as possible and should use common, everyday terms.
4. The questionnaires should be designed so that the answers to almost all questions are precoded.
5. The coding scheme for answers should be consistent across all questions.
6. The survey should include skip patterns, which indicate which questions are not to be asked, based on the answers given to the previous questions.

Once a questionnaire has been drafted by the person commissioned to work on the instrument, it should be presented to a team of experts for discussion. Everybody involved in the evaluation team (policy makers, researchers, data analysts, and data collectors) should be consulted about whether the questionnaire collects all the information desired in an appropriate fashion. Review by a team of experts is necessary but not sufficient, as intensive field testing is always primordial.

Direct observation of key outcomes is particularly important when the outcomes of interest may be hard to elicit truthfully from respondents. For instance, to avoid relying on self-reported data to measure outcomes related to crime or violence, some impact evaluations have embedded trained researchers in sample communities for them to observe subjects' behavior directly using ethnographic methods. Such direct observation can circumvent issues with self-reported behaviors, and can provide more accurate information when done well. Recent technological advances also allow direct measurements of a range of human behavior, and thus can help limit the use of self-reported data. Examples include direct observation of the timing and intensity of use of improved cookstoves, and direct measures of water quality, latrine use, and indoor temperature using electronic sensors.

Impact evaluations typically rely on a mix of traditional questionnaire-based surveys and other methods aimed at directly observing the outcomes of interest. For instance, in the context of impact evaluation of results-based financing in the health sector, a range of indicators are measured through complementary sources (Vermeersch, Rothenbühler, and Sturdy 2012). A health facility survey includes a facility assessment to measure the main characteristics of the facility, a health worker interview to measure health worker characteristics, and patient exit interviews to measure services delivered, as well as indicators of quality of care through a mix of vignettes and direct observation. A household survey includes household-level data on household and individual behavior, such as frequency of facility visits, care received, and health expenditures, as well as individual-level modules on female and child health. In addition to anthropometric measurement, biometric tests are collected to measure directly the prevalence of anemia, malaria, or HIV. Finally, community questionnaires capture community characteristics, services, infrastructure, access to markets, prices, and community-level shocks.

In addition to developing indicators and finding the most appropriate way to measure them, another key decision when collecting new data is the data collection technology to be used. Traditional data collection methods collect data based on paper, and later digitize that data, often through a double-blind data entry approach, which involves two separate agents digitizing the same information, before the data are compared to check for inaccuracies. Following recent technological advances, computer-assisted data collection tools have become prevalent. Data collection through applications installed on smartphones or tablets can speed up data processing, but also provide opportunities for real-time data quality checks and data validation. Box 16.4 discusses some of the pros and cons of electronic data collection.

Box 16.4: Some Pros and Cons of Electronic Data Collection

Computer-assisted personal interviewing (CAPI) provides an alternative to traditional pen-and-paper interviewing (PAPI). In CAPI, the survey is preloaded onto an electronic device, such as a tablet or smartphone. The interviewer reads the questions from the screen and enters the answers immediately into the device. Various software and applications have been developed for CAPI data collection. The pros and cons of CAPI must be carefully considered by the evaluation team.

Some pros:

- Electronic data collection can improve data quality. In a randomized experiment designed to compare CAPI and PAPI for a consumption survey in Tanzania, Caeyers, Chalmers, and De Weerdt (2012) found that data from paper surveys contained errors that were avoided in electronic surveys. The researchers discovered that the errors in the PAPI data were correlated with certain household characteristics, which can create bias in some data analysis.
- Electronic data collection programs can include automated consistency checks. Certain responses can trigger warning messages so that data entry errors are minimized and any issue is clarified with the respondent during the interview. For example, Fafchamps and others (2012) studied the benefits of consistency checks in a microenterprise survey in Ghana. They found that when consistency checks were introduced, the standard deviation of profit and sales data was lower. However, they also found that most of the time, a correction was not required: 85 percent to 97 percent of the time, respondents confirmed the original answer.
- Interviews can be shorter and easier to conduct. When CAPI is used, the flow of the questionnaire can be personalized to better guide interviewers through skip patterns and to minimize mistakes and omissions in the questionnaire. In a household survey in Tanzania, CAPI interviews were, on average, 10 percent shorter than similar questionnaires collected on paper, Caeyers, Chalmers, and De Weerdt (2012) found.
- Electronic data collection eliminates the need for manual reentry of data. This can reduce costs and speed up data processing.
- The use of technology can bring a range of indirect benefits. For example, by using tablets or smartphones, GPS coordinates can easily be collected, or photographs can be taken. Experimental variations in the survey content can also be introduced. With some software, parts of the interview can be recorded in order to facilitate quality and monitoring checks.

Some cons:

- The fixed costs tend to be higher for CAPI than PAPI, although the variable costs can be lower. The upfront cost of purchasing and programming electronic devices may be prohibitive for smaller impact evaluation budgets. Sufficient time is also needed up front to ensure proper programming and testing of the electronic questionnaires, which often comes after paper questionnaires have already been developed.

(continued)

Box 16.4: Some Pros and Cons of Electronic Data Collection *(continued)*

- Specific technical expertise is needed to program electronic questionnaires and set up processes to manage the flow of data collected electronically. In developing countries with low information technology capacity, this may be difficult to find. It is also more challenging to develop software for questionnaires that are not in English or a Romance language.
- Technological issues can disrupt data collection or hinder data consolidation in a secure location. Problems can arise during data collection when the electronic device has a small screen or an interface that is unfamiliar to interviewers. The risk of theft is also higher for electronic devices than paper surveys. Finally, the consolidation and synchronization of data in a secure location require clear protocols to minimize risk of data loss. Electronic transfers of data are convenient but require a minimum level of connectivity.

Sources: Caeyers, Chalmers, and De Weerdt 2012; Fafchamps and others 2012.

It is very important that the data collection instrument be piloted and field-tested extensively before it is finalized. Extensive *piloting* of the instrument will check its adequacy for the local context and its content, and any alternative formatting and phrasing options, as well as data collection protocols, including the technology. *Field-testing* the full data collection instrument in real-life conditions is critical for checking its length and for verifying that its format is sufficiently consistent and comprehensive to produce precise measures of all relevant information. Field-testing is an integral part of preparing the data collection instruments.

Conducting Fieldwork and Undertaking Quality Control

Even when you commission data collection, a clear understanding of all the steps involved in that process is crucial to help you ensure that the required *quality control mechanisms* and the right *incentives* are in place. The entity in charge of collecting data will need to coordinate the work of a large number of different actors, including enumerators, supervisors, field coordinators, and logistical support staff, in addition to a data entry team composed of programmers, supervisors, and the data entry operators. A clear *work plan* should be put in place to coordinate the work of all these teams, and the work plan is a key deliverable.

Before data collection begins, the work plan must include proper *training* for the data collection team. A complete *reference manual* should be prepared for training and used throughout fieldwork. Training is key to

ensuring that data are collected consistently by all involved. The training process is also a good opportunity to identify the best-performing enumerators and to conduct a last pilot of instruments and procedures under normal conditions. Once the sample has been drawn, the instruments have been designed and piloted, and the teams have been trained, the data collection can begin. It is good practice to ensure that the fieldwork plan has each survey team collect data on the same number of treatment and comparison units.

As discussed in chapter 15, proper sampling is essential to ensuring the quality of the sample. However, many *nonsampling errors* can occur while the data are being collected. In the context of an impact evaluation, a particular concern is that those errors may not be the same in the treatment and comparison groups.

Nonresponse arises when it becomes impossible to collect complete data for some sampled units. Because the actual samples used for analysis are restricted to those units for which data can be collected, units who choose not to respond to a survey may make the sample less representative and can create bias in the evaluation results. *Attrition* is a common form of nonresponse that occurs when some units drop from the sample between data collection rounds: for example, migrants may not be fully tracked.

Key Concept

Nonresponse arises when data are missing or incomplete for some sampled units. Nonresponse can create bias in the evaluation results.

Sample attrition due to nonresponse is particularly problematic in the context of impact evaluations because they may create differences between the treatment group and the comparison group. For example, *attrition* may be different in the two groups: if the data are being collected after the program has begun to be implemented, the response rate among treatment units can be higher than the rate among comparison units. That may happen because the comparison units are unhappy not to have been selected or are more likely to migrate. Nonresponses can also occur within the questionnaire itself, typically because some indicators are missing or the data are incomplete for a particular unit.

Measurement error is another type of problem that can generate bias if it is systematic. *Measurement error* is the difference between the value of a characteristic as provided by the respondent and the true (but unknown) value (Kasprzyk 2005). Such a difference can be traced to the way the questionnaire is worded or to the data collection method that is chosen, or it can occur because of the interviewers who are fielding the survey or the respondent who is giving the answers.

Key Concept

Best-practice impact evaluations aim to keep nonresponse and attrition as low as possible.

The quality of the impact evaluation depends directly on the quality of the data that are collected. *Quality standards* need to be made clear to all stakeholders in the data collection process; the standards should be particularly emphasized during the training of enumerators and in the reference manuals. For instance, detailed procedures to minimize nonresponse or

(if acceptable) to replace units in the sample are essential. The data collection agency must clearly understand the acceptable nonresponse and attrition rates. To provide a benchmark, many impact evaluations aim to keep nonresponse and attrition below 5 percent. The target will depend on the timing of the impact evaluation and the unit of analysis: attrition would be expected to be relatively lower for a survey occurring shortly after the baseline survey, and relatively higher for long-term impact evaluation tracking individuals many years later. Higher attrition rates would also be expected in very mobile populations. Survey respondents are sometimes compensated to minimize nonresponse, though the introduction of such compensation needs to be carefully considered. Sometimes, once all units to be tracked have been identified, a subsample of these units is randomly selected for very intensive tracking, which may include additional efforts or some form of compensation. In any case, the contract for the data collection agency must contain clear incentives, such as higher compensation if the nonresponse rate remains below an acceptable threshold.

Well-defined *quality assurance procedures* must be established for all stages of the data collection process, including the design of the sampling procedure and questionnaire, the preparation stages, data collection, data entry, and data cleaning and storage.

Quality checks during the fieldwork should be given a very high priority to minimize errors for each unit. Clear procedures must exist for revisiting units that have provided no information or incomplete information. Multiple filters should be introduced in the quality control process: for instance, by having enumerators, supervisors, and if necessary, field coordinators revisit the nonresponse units to verify their status. The questionnaires from nonresponse interviews should still be clearly coded and recorded. Once the data have been completely digitized, the nonresponse rates can be summarized and all sampled units fully accounted for.

Quality checks should also be made on any incomplete data for a particular surveyed unit. Again, the quality control process should include multiple filters. The enumerator is responsible for checking the data immediately after they have been collected. The supervisor and the field coordinator should perform random checks at a later stage.

Quality checks for measurement errors are more difficult but are crucial for assessing whether information has been collected accurately. Consistency checks can be built into the questionnaire. In addition, supervisors or quality controllers need to conduct *spot checks* by participating in interviews to ensure that the enumerators collect data in accordance with the established quality standards. *Back-checks* or quality audits can be undertaken among a subsample of the impact evaluation survey to ensure that the data collected are accurate. This is sometimes done by having a

quality controller collect a subset of the questionnaire with a respondent, and comparing the response with those previously obtained by an enumerator with the same respondent.

Field coordinators or members of the evaluation team should also contribute to quality checks to minimize potential conflicts of interest within the survey firm. You may also consider contracting with an external agency to audit the quality of the data collection activities. Doing that can significantly limit the range of problems that can arise as a result of lack of supervision of the data collection team or insufficient quality control procedures.

Ultimately, it is critical that all steps involved in checking quality are requested explicitly in the terms of reference when commissioning data collection.

Processing and Storing the Data

Data processing and validation is an integral part of the collection of new survey data. It includes the steps to digitize information in paper-and-pencil surveys, as well as the steps to validate data for both paper-and-pencil surveys and electronic data collection using laptop computers, smartphones, tablets, or other devices. When working with paper-and-pencil surveys, *a data entry program* must be developed and a system put in place to manage the flow of data to be digitized. Norms and procedures must be established, and data entry operators must be carefully trained to guarantee that data entry is consistent. As much as possible, data entry should be integrated into data collection operations (including during the pilot-testing phase), so that any problems with the data collected can be promptly identified and verified in the field. Overall, the quality benchmark for the data entry process should be that the raw physical data are exactly replicated in the digitized version, with no modifications made to them while they are being entered. To minimize data entry errors, a *double-blind data entry* procedure can be used to identify and correct for any remaining errors. A computer assisted field entry (CAFE) approach can be used, which collects data in a paper-and-pencil survey, and then digitizes it in the field and immediately validates it to identify errors and inconsistencies.

For both paper-and-pencil surveys and surveys relying on electronic data collection, programs can be developed to perform automatic checks for nonsampling errors (both item nonresponse and inconsistencies) that may occur in the field and to validate data. If the validation process is integrated into the fieldwork procedures, incomplete or inconsistent data can be referred back to the fieldworkers for on-site verification.

This kind of integration is not without challenges for the organizational flow of fieldwork operations, but it can yield substantial gains in quality, diminish measurement error, and increase the statistical power of the impact evaluation. The possibility of using such an integrated approach should be considered explicitly when data collection is being planned. The use of new technologies can facilitate those quality checks.

As discussed, data collection comprises a set of operations whose complexity should not be underestimated. Box 16.5 discusses how the data collection process for the evaluation of the Atención a Crisis pilots

Box 16.5: Data Collection for the Evaluation of the Atención a Crisis Pilots in Nicaragua

In 2005, the Nicaraguan government launched the Atención a Crisis pilot program. A study was set up to evaluate the impact of combining a conditional cash transfer (CCT) program with productive transfers, such as grants for investment in nonagricultural activities or vocational training. The Atención a Crisis pilot was implemented by the ministry of the family, with support from the World Bank.

A randomized assignment in two stages was used for the evaluation. First, 106 target communities were randomly assigned to either the comparison group or the treatment group. Second, within treatment communities, eligible households were randomly assigned one of three benefit packages: a conditional cash transfer; the CCT plus a scholarship that allowed one of the household members to choose among a number of vocational training courses; and the CCT plus a productive investment grant to encourage recipients to start a small nonagricultural activity, with the goal of creating assets and diversifying income (Macours, Premand, and Vakis 2012).

A baseline survey was collected in 2005, a first follow-up survey occurred in 2006, and a second follow-up survey was conducted in 2008, two years after the intervention ended. Rigorous quality checks were put in place at all stages of the data collection process. First, questionnaires were thoroughly field-tested, and enumerators were trained in both class and field conditions. Second, field supervision was set up so that all questionnaires were revised multiple times by enumerators, supervisors, field coordinators, and other reviewers. Third, a double-blind data entry system was used, together with a comprehensive quality-check program that could identify incomplete or inconsistent questionnaires. Questionnaires with missing information in certain questions or inconsistencies were systematically sent back to the field for verification. These procedures and requirements were explicitly specified in the terms of reference of the data collection firm.

(continued)

Box 16.5: Data Collection for the Evaluation of the Atención a Crisis Pilots in Nicaragua *(continued)*

In addition, detailed tracking procedures were put in place to minimize attrition. At the start, a full census of households residing in the treatment and control communities in 2008 was undertaken in close collaboration with community leaders. Because migration within the country was common the survey firm was given incentives to track individual migrants throughout the country. As a result, only 2 percent of the original 4,359 households could not be interviewed in 2009. The survey firm was also commissioned to track all individuals from the households surveyed in 2005. Again, only 2 percent of the individuals to whom program transfers were targeted could not be tracked (another 2 percent had died). Attrition was 3 percent for all children of households surveyed in 2005 and 5 percent for all individuals in households surveyed in 2005.

Attrition and nonresponse rates provide a good indicator of survey quality. Reaching very low attrition rates required intense efforts by the data collection firm, as well as explicit incentives. The per unit cost of a tracked household or individual is also much higher. In addition, thorough quality checks added costs and increased data collection time. Still, in the context of the Atención a Crisis pilot, the sample remained representative at both the household and the individual levels three to four years after the baseline, measurement error was minimized, and the reliability of the evaluation data was ensured. As a result, the long-term impacts of the Atención a Crisis pilots could be convincingly analyzed.

Source: Macours, Premand, and Vakis 2012.

in Nicaragua yielded high-quality data with very low attrition and item nonresponse and few measurement and processing errors. Such high-quality data can be obtained only when data quality procedures and proper incentives are put in place at the moment of commissioning data collection.

At the end of the data collection process, the data set should be delivered with detailed documentation, including a complete codebook and data dictionary, and stored in a secure location (see box 16.6). If the data are being collected for an impact evaluation, then the data set should also include complementary information on treatment status and program participation. A complete set of documentation will speed up the analysis of the impact evaluation data, help produce results that can be used for policy making in a timely fashion, and facilitate information sharing and potential replication.

Box 16.6: Guidelines for Data Documentation and Storage

The key guideline in data documentation is to keep a record of all impact evaluation data. This includes data collection protocols, questionnaires, training manuals, and the like. The World Bank, Inter-American Development Bank, and the Millennium Challenge Corporation, among others, have open data initiatives where this data is made publicly available via a data catalog.

Data storage can be broken up into three categories: microdata, macrodata, and identification (ID) control files.

- *Microdata* are data at the level of the unit of observation that are made anonymous and do not include any information identifying the individuals. Relevant identifying variables have been anonymized with IDs, which are linked only to respondent information in ID control files.
- *ID control files* contain the full information before it is made anonymous. They should be saved only in a secure server and never included in a data catalogue.
- *Macrodata* include all supporting documents that are relevant to the interpretation of the microdata, such as a data dictionary, codebook, description of the study design, and questionnaires.

Cataloguing macrodata and microdata helps protect the security of the data and also follows international standards on data storage. Central data catalogues are much less vulnerable to malfunction or hacking than a computer hard drive or portable storage device. Within certain data catalogues, the data can be password-protected for a period of time before becoming publicly available.

Additional Resources

- For accompanying material to the book and hyperlinks to additional resources, please see the Impact Evaluation in Practice website (http://www.worldbank.org/ieinpractice).
- For a guide to questionnaire design, see the module on "Applied Fieldwork Techniques" in the applied impact evaluation methods course at University of California (http://aie.cega.org).
- For blog posts about data collection, see the curated list on the World Bank Development Impact blog (http://blogs.worldbank.org/impactevaluations).
- For more information on data collection, see the following:
  - Fink, Arlene G., and Jacqueline Kosecoff. 2008. *How to Conduct Surveys: A Step by Step Guide*, fourth edition. London: Sage.
  - Iarossi, Giuseppe. 2006. *The Power of Survey Design: A User's Guide for Managing Surveys, Interpreting Results, and Influencing Respondents.* Washington, DC: World Bank.
  - Leeuw, Edith, Joop Hox, and Don Dillman. 2008. *International Handbook of Survey Methodology.* New York: Taylor & Francis Group.

- For more on data collection activities and data quality oversight, see the World Bank Impact Evaluation Toolkit, Module 5 on Data Collection (http://www.worldbank.org/health/impactevaluationtoolkit). The module includes several examples of survey progress reports, field manuals, and training programs for households and health facilities.
- For a variety of materials for guidance on preparing a survey, see the Inter-American Development Bank Evaluation hub (http://www.iadb.org/evaluationhub). In the Data Collection section, you can download
  - A questionnaire designer manual
  - A data entry manual
  - Consent forms, sample questionnaires, data entry programs, and fieldwork manuals for several different types of surveys, including surveys for households, communities, health facilities, schools, and farmers
  - Links to further examples of survey questions and questionnaires
  - Links to guidelines for quality data collection
  - Links to tools available on the International Household Survey Network (IHSN) website for data storage and management.
- For more on why data documentation is important, how it can be done, and who within the evaluation team is responsible for it, see the World Bank Impact Evaluation Toolkit, Module 6 on Data Storage (http://www.worldbank.org/health/impactevaluationtoolkit).

Notes

1. See references in Grosh and Glewwe (2000) and UN (2005). See also Muñoz (2005); Iarossi (2006); Fink and Kosecoff (2008); and Leeuw, Hox, and Dillman (2008), which provide a wealth of practical guidance for data collection.
2. See McKenzie and Rosenzweig (2012) for an overview of recent advances.
3. For examples of such experiments, see McKenzie and Rosenzweig (2012) on general issues; Beegle, Carletto, and Himelein (2012) on agricultural data; Beegle and others (2012) on measuring household consumption; and Bardasi and others (2011) on labor data.
4. For examples of innovations in measuring outcomes, see Holla (2013); Das and Hammer (2007); and Planas and others (2015).

References

Baird, S., and B. Özler. 2012. "Examining the Reliability of Self-reported Data on School Participation." *Journal of Development Economics* 98 (1): 89–93.

Bardasi, E., K. Beegle, A. Dillon, A., and P. Serneels. 2011. "Do Labor Statistics Depend on How and to Whom the Questions Are Asked? Results from a Survey Experiment in Tanzania." *The World Bank Economic Review* 25 (3): 418–47.

Beegle, K., C. Carletto, and K. Himelein. 2012. "Reliability of Recall in Agricultural Data." *Journal of Development Economics* 98 (1): 34–41.

Beegle, K., J. De Weerdt, J. Friedman, and J. Gibson. 2012. "Methods of Household Consumption Measurement through Surveys: Experimental Results from Tanzania." *Journal of Development Economics* 98 (1): 3–18.

Caeyers, Bet, Neil Chalmers, and Joachim De Weerdt. 2012. "Improving Consumption Measurement and Other Survey Data through CAPI: Evidence from a Randomized Experiment." *Journal of Development Economics* 98 (1): 19–33.

Chetty, R., J. N. Friedman, and E. Saez. 2013. "Using Differences in Knowledge across Neighborhoods to Uncover the Impacts of the EITC on Earnings." *American Economic Review* 103 (7): 2683–721.

Das, J., and J. Hammer. 2007. "Money for Nothing: The Dire Straits of Medical Practice in Delhi, India." *Journal of Development Economics* 83 (1): 1–36.

Fafchamps, Marcel, David McKenzie, Simon Quinn, and Christopher Woodruff. 2012. "Using PDA Consistency Checks to Increase the Precision of Profits and Sales Measurement in Panels." *Journal of Development Economics* 98 (1): 51–57.

Ferraz, C., and F. Finan. 2008. "Exposing Corrupt Politicians: The Effects of Brazil's Publicly Released Audits on Electoral Outcomes." *The Quarterly Journal of Economics* 123 (2): 703–45.

Fink, A. G., and J. Kosecoff. 2008. *How to Conduct Surveys: A Step by Step Guide*, fourth edition. London: Sage.

Galiani, S., P. Gertler, and E. Schargrodsky. 2005. "Water for Life: The Impact of the Privatization of Water Services on Child Mortality." *Journal of Political Economy* 113 (1): 83–120.

Galiani, Sebastian, and Patrick McEwan. 2013. "The Heterogeneous Impact of Conditional Cash Transfers." *Journal of Public Economics* 103: 85–96.

Gertler, Paul, Paula Giovagnoli, and Sebastian Martinez. 2014. "Rewarding Provider Performance to Enable a Healthy Start to Life: Evidence from Argentina's Plan Nacer." Policy Research Working Paper 6884, World Bank, Washington, DC.

Glewwe, Paul. 2005. "An Overview of Questionnaire Design for Household Surveys in Developing Countries." In *Household Sample Surveys in Developing and Transition Countries*. New York: United Nations.

Glewwe, Paul, and Pedro Olinto. 2004. "Evaluating the Impact of Conditional Cash Transfers on Schooling: An Experimental Analysis of Honduras' PRAF Program." Final report. University of Minnesota and IFPRI-FCND.

Grosh, Margaret, and Paul Glewwe, eds. 2000. *Designing Household Survey Questionnaires for Developing Countries: Lessons from 15 Years of the Living Standards Measurement Study*. Washington, DC: World Bank.

Holla, Alaka. 2013. "Measuring the Quality of Health Care in Clinics." World Bank, Washington, DC. http://www.globalhealthlearning.org/sites/default/files/page-files/Measuring%20Quality%20of%20Health%20Care_020313.pdf.

Iarossi, G. 2006. *The Power of Survey Design: A User's Guide for Managing Surveys, Interpreting Results, and Influencing Respondents*. Washington, DC: World Bank.

Kasprzyk, Daniel. 2005. "Measurement Error in Household Surveys: Sources and Measurement." In *Household Sample Surveys in Developing and Transition Countries*. New York: United Nations.

Leeuw, E., J. Hox, and D. Dillman. 2008. *International Handbook of Survey Methodology.* New York: Taylor & Francis Group.

Macours, Karen, Patrick Premand, and Renos Vakis. 2012. "Transfers, Diversification and Household Risk Strategies: Experimental Evidence with Implications for Climate Change Adaptation." Policy Research Working Paper 6053, World Bank, Washington, DC.

McKenzie, David, and Mark Rosenzweig. 2012. "Symposium on Measurement and Survey Design." *Journal of Development Economics* 98 (May 1): 1–148.

Morris, Saul S., Rafael Flores, Pedro Olinto, and Juan Manuel Medina. 2004. "Monetary Incentives in Primary Health Care and Effects on Use and Coverage of Preventive Health Care Interventions in Rural Honduras: Cluster Randomized Trial." *Lancet* 364: 2030–37.

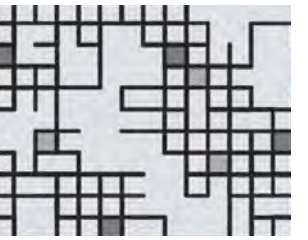
Muñoz, Juan. 2005. "A Guide for Data Management of Household Surveys." In *Household Sample Surveys in Developing and Transition Countries,* chapter 15. New York: United Nations.

Planas, M-E, P. J. García, M. Bustelo, C. P. Carcamo, S. Martinez, H. Nopo, J. Rodriquez, M-F Merino, and A. Morrison. 2015. "Effects of Ethnic Attributes on the Quality of Family Planning Services in Lima, Peru: A Randomized Crossover Trial." *PLoS ONE* 10 (2): e0115274.

Pradhan, M., and L. B. Rawlings. 2002. "The Impact and Targeting of Social Infrastructure Investments: Lessons from the Nicaraguan Social Fund." *World Bank Economic Review* 16 (2): 275–95.

UN (United Nations). 2005. *Household Sample Surveys in Developing and Transition Countries.* New York: United Nations.

Vermeersch, Christel, Elisa Rothenbühler, and Jennifer Sturdy. 2012. *Impact Evaluation Toolkit: Measuring the Impact of Results-Based Financing on Maternal and Child Health*. World Bank, Washington, DC. http://www.worldbank.org/health/impactevaluationtoolkit.



CHAPTER 17

Conclusion

Impact Evaluations: Worthwhile but Complex Exercises

Impact evaluation is about generating evidence about which programs work, which do not, and how to improve them to achieve better development outcomes. That can be done in a classic impact evaluation framework, comparing outcomes between treatment and comparison groups. Impact evaluations can also be conducted to explore implementation alternatives within a program, to test innovations, or to look across programs to assess comparative performance.

We argue that impact evaluations are a worthwhile investment for many programs. Coupled with monitoring and other forms of evaluation, they enhance the understanding of the effectiveness of particular policies; they contribute to improved accountability for program managers, governments, funders, and the public; they inform decisions about how to allocate scarce development resources more efficiently; and they add to the global store of knowledge about what works and what does not in the field of development.

Checklist: Core Elements of a Well-Designed Impact Evaluation

Impact evaluations are complex undertakings with many moving parts. The following checklist highlights the core elements of a well-designed impact evaluation:

- ✓ A concrete and relevant policy question—grounded in a theory of change—that can be answered with an impact evaluation
- ✓ A robust methodology, derived from the operational rules of the program, to estimate a counterfactual that shows the causal relationship between the program and outcomes of interest
- ✓ A well-formed evaluation team that functions as a partnership between a policy team and a research team
- ✓ A respect for ethical standards and consideration of human subjects in the design and implementation of the evaluation and related data collection, as well as attention to open science principles to ensure transparency
- ✓ A sample with sufficient statistical power to allow policy-relevant impacts to be detected
- ✓ A methodology and sample that provide results generalizable for the population of interest
- ✓ High-quality data that provide the appropriate information required for the impact evaluation, including data for the treatment and comparison groups, data at baseline and follow-up, and information on program implementation and costs
- ✓ An engagement strategy to inform policy dialogue through the implementation of the impact evaluation, as well as an impact evaluation report and associated policy briefs disseminated to key audiences in a timely manner.

Checklist: Tips to Mitigate Common Risks in Conducting an Impact Evaluation

We also highlight some tips that can help mitigate common risks inherent in the process of conducting an impact evaluation:

- ✓ Impact evaluations are best designed early in the project cycle, ideally as part of the program design, but at least before the program to be evaluated is implemented. Early planning allows for a prospective evaluation

design based on the best available methodology and will provide the time necessary to plan and implement baseline data collection in evaluation areas before the program starts.

- ✓ Impact evaluation results should be informed by complementary process evaluation and monitoring data that give a clear picture of program implementation. When programs succeed, it is important to understand why. When programs fail, it is important to distinguish between a poorly implemented program and a flawed program design.
- ✓ Baseline data should be collected, and a backup methodology should be built into your impact evaluation design. If the original evaluation design is invalidated—for example, because the original comparison group receives program benefits—having a backup plan can help you avoid having to throw out the evaluation altogether.
- ✓ Common identifiers should be maintained among different data sources for your units of observation so that they can be easily linked during the analysis. For example, a particular household should have the same identifier in the monitoring systems and in baseline and follow-up impact evaluation surveys.
- ✓ Impact evaluations are useful for learning about how programs work and for testing program alternatives, even for large ongoing programs. Well-designed impact evaluations can help test innovations or provide insights on the relative effectiveness of various goods and services delivered as a bundle in existing programs. Embedding an additional program innovation as a small pilot in the context of a larger evaluation can leverage the evaluation to produce valuable information for future decision making.
- ✓ Impact evaluations should be thought of as another component of a program's operation and should be adequately staffed and budgeted with the required technical and financial resources. Be realistic about the costs and complexity of carrying out an impact evaluation. The process of designing an evaluation and collecting a baseline from scratch can typically take a year or more. Once the program starts, the treatment group needs a sufficient period of exposure to the intervention to affect outcomes. Depending on the program, that can take anywhere from one year to five years, or more for long-term outcomes. Collecting one or more follow-up surveys, conducting the analysis, and disseminating the results will also involve substantial effort over a number of months and years. Altogether, a complete impact evaluation cycle from start to finish typically takes at least three to four years of intensive work and engagement. Adequate financial and technical resources are necessary at each step of the way.

Ultimately, individual impact evaluations provide concrete answers to specific policy questions. Although these answers provide information that is customized for the specific entity commissioning and paying for the evaluation, they also provide information that is of value to others around the world who can learn and make decisions based on the evidence. For example, cash transfer programs in Africa, Asia, and Europe have drawn lessons from the original evaluations of Colombia's Familias en Acción, Mexico's Progresa, and other Latin American conditional cash transfer programs. In this way, impact evaluations are partly a global public good. Evidence generated through one impact evaluation adds to global knowledge on that subject. This knowledge base can then inform policy decisions in other countries and contexts as well, with appropriate attention to external validity. The international community has been moving rapidly toward scaling up support for rigorous evaluation.

At the country level, more sophisticated and demanding governments are looking to demonstrate results and to be more accountable to their core constituencies. Increasingly, evaluations are being conducted by national and subnational line ministries and government bodies set up to lead a national evaluation agenda, such as the National Council for Evaluation of Social Development Policies in Mexico and the Department of Performance Monitoring and Evaluation in South Africa. Evidence from impact evaluations is also being used to inform budgetary allocations made by congress and parliament at the national level. In systems where programs are judged based on hard evidence and final outcomes, programs with a strong evidence base to defend positive results will be able to thrive, while programs lacking such proof will find it more difficult to sustain funding.

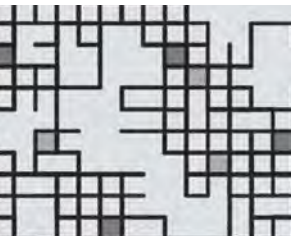
Multilateral institutions such as the World Bank and the Inter-American Development Bank, as well as national development agencies, donor governments, and philanthropic institutions, are also demanding more and better evidence on the effective use of development resources. Such evidence is required for accountability to those lending or donating the money, as well as for decision making about where best to allocate scarce development resources.

A growing number of institutions dedicated primarily to the production of high-quality impact evaluations are expanding, including ones from the academic arena, such as the Poverty Action Lab, Innovations for Poverty Action, and the Center for Effective Global Action, and independent agencies that support impact evaluations, such as the International Initiative for Impact Evaluation (3ie). A number of associations bring together groups of evaluation practitioners and researchers and policy makers interested in the topic, including the Network of Networks on Impact Evaluation and regional associations such as the African Evaluation Association and the

Latin American and Caribbean Economics Association Impact Evaluation Network. All these efforts reflect the increasing importance of impact evaluation in international development policy.

Given this growth in impact evaluation, being conversant in the language of impact evaluation is an increasingly indispensable skill for any development practitioner—whether you run evaluations for a living, contract impact evaluations, or use the results of impact evaluations for decision making. Rigorous evidence of the type generated through impact evaluations can be one of the drivers of development policy dialogue, providing the basis to support or oppose investments in development programs and policies. Evidence from impact evaluations allows policy makers and project managers to make informed decisions on how to achieve outcomes more cost-effectively. Equipped with the evidence from an impact evaluation, the policy team has the job of closing the loop by feeding those results into the decision-making process. This type of evidence can inform debates, opinions, and ultimately, the human and monetary resource allocation decisions of governments, multilateral institutions, and donors.

Evidence-based policy making is fundamentally about informing program design and better allocating budgets to expand cost-effective programs, curtail ineffective ones, and introduce improvements to program designs based on the best available evidence. Impact evaluation is not a purely academic undertaking. Impact evaluations are driven by the need for answers to policy questions that affect people's daily lives. Decisions on how best to spend scarce resources on antipoverty programs, transport, energy, health, education, safety nets, microcredit, agriculture, and myriad other development initiatives have the potential to improve the welfare of people across the globe. It is vital that those decisions be made using the most rigorous evidence possible.



GLOSSARY

Italicized terms within the definitions are also defined elsewhere in the glossary.

Activity. Actions taken or work performed through which *inputs*, such as funds, technical assistance, and other types of resources, are mobilized to produce specific *outputs*, such as money spent, textbooks distributed, or number of participants enrolled in an employment program.

Administrative data. Data routinely collected by public or private agencies as part of program administration, usually at a regular frequency and often at the point of service delivery, including services delivered, costs, and program participation. *Monitoring* data are a type of *administrative data.*

Alternative hypothesis. The *hypothesis* that the *null hypothesis* is false. In an *impact evaluation*, the alternative hypothesis is usually the hypothesis that the *intervention* has an impact on *outcome*s.

Attrition. Attrition occurs when some units drop out from the *sample* between one round of data collection and another, for example, when people move and can't be located. Attrition is a case of *unit nonresponse*. Attrition can create *bias* in the impact estimate.

Average treatment effect (ATE). The impact of the program under the assumption of full *compliance;* that is, all *units* that have been assigned to a program actually enroll in it, and none of the comparison units receive the program.

Baseline. The state before the *intervention*, against which progress can be assessed or comparisons made. Baseline data are collected before a program or policy is implemented to assess the *before* state. The availability of baseline data is important to document balance in preprogram characteristics between treatment and comparison groups. Baseline data are required for some *quasi-experimental* designs.

Before-and-after comparison. Also known as *pre-post comparison* or *reflexive comparison*. This strategy tracks changes in *outcomes* for program beneficiaries over time, using measurements before and after the program or policy is implemented, without using a *comparison group*.

Bias. In *impact evaluation*, bias is the difference between the impact that is calculated and the true impact of the program.

Causal effect. See *impact*.

Census. A complete enumeration of a population. Census data cover all units in the population. Contrast with *sample*.

Cluster. Units that are grouped and may share similar characteristics. For example, children who attend the same school would belong to a cluster because they share the same school facilities and teachers and live in the same neighborhood.

Clustered sample. A *sample* composed of clusters.

Comparison group. Also known as a *control group*. A valid comparison group will have the same characteristics on average as the group of beneficiaries of the program (*treatment group*), except for the fact that the units in the comparison group do not benefit from the program. Comparison groups are used to estimate the *counterfactual*.

Compliance. Compliance occurs when *units* adhere to their assignment to the *treatment group* or *comparison group*.

Context equilibrium effects. *Spillovers* that happen when an *intervention* affects the behavioral or social norms within a given context, such as a treated locality.

Control group. Also known as a *comparison group* (see definition).

Correlation. A statistical measure that indicates the extent to which two or more *variables* fluctuate together.

Cost-benefit analysis. Estimates the total expected benefits of a program, compared with its total expected costs. It seeks to quantify all of the costs and benefits of a program in monetary terms and assesses whether benefits outweigh costs.

Cost-effectiveness analysis. Compares the relative cost of two or more programs or program alternatives in terms of reaching a common *outcome*, such as agricultural yields or student test scores.

Counterfactual. What the *outcome* (*Y*) would have been for program participants if they had not participated in the program (*P*). By definition, the counterfactual cannot be observed. Therefore, it must be estimated using a *comparison group*.

Coverage bias. Occurs when a *sampling frame* does not exactly coincide with the *population of interest*.

Crossover design. Also called a cross-cutting design. This is when there is *randomized assignment* with two or more interventions, allowing the impact of individual and combined interventions to be estimated.

Data mining. The practice of manipulating the data in search of particular results.

Dependent variable. Usually the *outcome* variable. The variable to be explained, as opposed to *explanatory variables*.

Difference-in-differences. Also known as *double difference* or *DD*. Difference-in-differences compares the changes in *outcomes* over time between the *treatment group* and the *comparison group*. This eliminates any differences between these groups that are constant over time.

Effect size. The magnitude of the change in an *outcome* that is caused by an *intervention*.

Effectiveness study. Assesses whether a program works under normal conditions at scale. When properly designed and implemented, results from these studies can be more generalizable than *efficacy studies*.

Efficacy study. Assesses whether a program can work under ideal conditions. These studies are carried out under very specific circumstances, for example, with heavy technical involvement from researchers during implementation of the program. They are often undertaken to test the viability of a new program. Their results may not be not generalizable beyond the scope of the evaluation.

Eligibility index. Also known as the *forcing variable*. A *variable* that ranks the *population of interest* along a continuum and has a threshold or cutoff score that determines who is eligible and who is not.

Enrolled-and-nonenrolled comparisons. Also known as *self-selected comparisons*. This strategy compares the outcomes of *units* that choose to enroll and units that choose not to enroll in a program.

Estimator. In statistics, an estimator is a rule that is used to estimate an unknown population characteristic (technically known as a *parameter*) from the data; an estimate is the result from the actual application of the rule to a particular sample of data.

Evaluation. A periodic, objective assessment of a planned, ongoing, or completed project, program, or policy. Evaluations are used to answer specific questions, often related to design, implementation, or results.

Evaluation team. The team that conducts the *evaluation*. It is essentially a partnership between two groups: a team of policy makers and program managers (the policy team) and a team of researchers (the research team).

Ex ante simulations. *Evaluations* that use available data to simulate the expected effects of a program or policy reform on *outcomes* of interest.

Explanatory variable. Also known as the *independent* variable**.** A *variable* that is used on the right-hand side of a regression to help explain the *dependent variable* on the left-hand side of the regression.

External validity. An *evaluation* is externally valid if the evaluation *sample* accurately represents the population of interest of eligible *units*. The results of the evaluation can

then be generalized to the population of eligible units. Statistically, for an *impact evaluation* to be externally valid, the evaluation *sample* must be representative of the *population of interest*. Also see *internal validity*.

Follow-up survey. Also known as a *postintervention survey*. A survey that is fielded after the program has started, once the beneficiaries have benefited from it for some time. An *impact evaluation* can include several follow-up surveys, which are sometimes referred as *midline* and *endline surveys*.

General equilibrium effects. *Spillovers* that happen when *interventions* affect the supply and demand for goods or services, and thereby change the market price for those goods or services.

Generalizability. The extent to which results from an *evaluation* carried out locally will hold true in other settings and among other population groups.

Hawthorne effect. Occurs when the mere fact that units are being observed makes them behave differently.

Hypothesis. A proposed explanation for an observable phenomenon. See also, *null hypothesis* and *alternative hypothesis.*

Impact. Also known as *causal effect*. In the context of *impact evaluations*, an impact is a change in outcomes that is directly attributable to a program, program modality, or design innovation.

Impact evaluation. An *evaluation* that makes a causal link between a program or *intervention* and a set of *outcome*s. An impact evaluation answers the question: What is the *impact* (or causal effect) of a program on an outcome of interest.

Imperfect compliance. The discrepancy between assigned treatment status and actual treatment status. Imperfect compliance happens when some units assigned to the *comparison group* participate in the program, or some units assigned to the *treatment group* do not.

Indicator. A *variable* that measures a phenomenon of interest to the evaluation team. The phenomenon can be an *input*, an *output*, an *outcome*, a characteristic, or an attribute. Also see *SMART*.

Informed consent. One of the cornerstones of protecting the rights of human subjects. In the case of *impact evaluations*, it requires that respondents have a clear understanding of the purpose, procedures, risks, and benefits of the data collection that they are asked to participate in.

Inputs. The financial, human, and material resources used for the *intervention.*

Institutional Review Board (IRB). A committee that has been designated to review, approve, and monitor research involving human subjects. Also known as an *independent ethics committee* (IEC) or *ethical review board* (ERB).

Instrumental variable. Also known as *instrument*. The instrumental variable method relies on some external source of variation or IV to determine treatment status.

The IV influences the likelihood of participating in a program, but it is outside of the participant's control and is unrelated to the participant's characteristics.

Intention-to-treat (ITT). ITT estimates measure the difference in outcomes between the units assigned to the *treatment group* and the units assigned to the *comparison group,* irrespective of whether the units assigned to either group actually receive the treatment.

Internal validity. An *evaluation* is internally valid if it provides an accurate estimate of the *counterfactual* through a valid *comparison group*.

Intervention. In the context of impact evaluation, this is the project, program, design innovation, or policy to be evaluated. Also known as the *treatment*.

Intra-cluster correlation. Also known as *intra-class correlation.* This is the degree of similarity in *outcome*s or characteristics among units within preexisting groups or *clusters,* relative to units in other clusters. For example, children who attend the same school would typically be more similar or correlated in terms of their area of residence or socioeconomic background, relative to children who don't attend this school.

Item nonresponse. Occurs when data are incomplete for some sampled *units*.

John Henry effect. The John Henry effect happens when comparison units work harder to compensate for not being offered a treatment. When we compare treated units with those harder-working comparison units, the estimate of the impact of the program will be *biased:* that is, we will estimate a smaller impact of the program than the true impact that we would find if the comparison units did not make the additional effort.

Lack of common support. When using the *matching* method, lack of common support is a lack of overlap between the *propensity scores* of the treatment or enrolled group and those of the pool of nonenrolled.

Local average treatment effect (LATE). The *impact* of the program estimated for a specific subset of the population, such as *units* that comply with their assignment to the treatment or comparison group in the presence of *imperfect compliance,* or around the eligibility cutoff score when applying a *regression discontinuity design.* Thus the LATE provides only a local estimate of the program impact and should not be generalized to the entire population.

Matching. A nonexperimental *impact evaluation* method that uses large data sets and statistical techniques to construct the best possible *comparison group* for a given *treatment group* based on observed characteristics.

Mechanism experiment. An *impact evaluation* that tests a particular causal mechanism within the *theory of change* of a program, rather than testing the causal effect (*impact*) of the program as a whole.

Minimum detectable effect. The minimum detectable effect is an input for *power calculations;* that is, it provides the effect size that an *impact evaluation* is designed to estimate for a given level of *significance* and *power*. Evaluation

samples need to be large enough to detect a policy-relevant minimum detectable effect with sufficient power. The minimum detectable effect is set by considering the change in *outcome*s that would justify the investment in an *intervention*.

Mixed methods. An analytical approach that combines quantitative and qualitative data.

Monitoring. The continuous process of collecting and analyzing information to assess how well a project, program, or policy is performing. Monitoring usually tracks *input*s, *activities*, and *output*s, though occasionally it also includes *outcome*s. Monitoring is used to inform day-to-day management and decisions. It can also be used to track performance against expected results, make comparisons across programs, and analyze trends over time.

Monitoring data. Data from program *monitoring* that provide essential information about the delivery of an *intervention*, including who the beneficiaries are and which program benefits or *outputs* they may have received. Monitoring data are a type of *administrative data*.

Nonresponse. Occurs when data are missing or incomplete for some sampled units. *Unit nonresponse* arises when no information is available for some *sample* units: that is, when the actual sample is different from the planned sample. One form of unit nonresponse is *attrition*. *Item nonresponse* occurs when data are incomplete for some sampled units at a point in time. Nonresponse may cause *bias* in *evaluation* results if it is associated with treatment status.

Null hypothesis. A *hypothesis* that might be falsified on the basis of observed data. The null hypothesis typically proposes a general or default position. In *impact evaluation*, the null hypothesis is usually that the program does not have an *impact*; that is, that the difference between outcomes in the *treatment group* and the *comparison group* is zero.

Open science. A movement that aims to make research methods more transparent, including through trial registration, use of preanalysis plans, data documentation, and registration.

Outcome. A result of interest that is measured at the level of program beneficiaries. Outcomes are results to be achieved once the beneficiary population uses the project outputs. Outcomes are not directly under the control of a program-implementing agency: they are affected both by the implementation of a program (the *activities* and *outputs* it delivers) and by behavioral responses from beneficiaries exposed to that program (the use that beneficiaries make of the benefits they are exposed to). An outcome can be intermediate or final (long term). Final outcomes are more distant outcomes. The distance can be interpreted in terms of time (it takes a longer period of time to get to the outcome) or in terms of causality (many causal links are needed to reach the outcome and multiple factors influence it).

Output. The tangible products, goods, and services that are produced (supplied) directly by a program's *activities*. The delivery of outputs is directly under the control

of the program-implementing agency. The use of outputs by beneficiaries contributes to changes in *outcomes*.

Placebo test. Falsification test used to assess whether the assumptions behind a method hold. For instance, when applying the *difference-in-differences* method, a placebo test can be implemented by using a fake treatment group or fake outcome: that is, a group or outcome that you know was not affected by the program. Placebo tests cannot confirm that the assumptions hold but can highlight cases when the assumptions do not hold.

Population of interest. A comprehensive group of all *units* (such as individuals, households, firms, facilities) that are eligible to receive an intervention or treatment, and for which an *impact evaluation* seeks to estimate program *impacts*.

Power (or *statistical power*). The probability that an impact evaluation will detect an impact (that is, a difference between the *treatment group* and *comparison group*) when in fact one exists. The power is equal to 1 minus the probability of a *type II error*, ranging from 0 to 1. Common levels of power are 0.8 and 0.9. High levels of power are more conservative, meaning that there is a low likelihood of not detecting real program impacts.

Power calculations. Calculations to determine how large a *sample* size is required for an *impact evaluation* to precisely estimate the impact of a program: that is, the smallest sample that will allow us to detect the *minimum detectable effect*. Power calculations also depend on parameters such as *power* (or the likelihood of *type II error*), *significance level*, mean, variance, and *intra-cluster correlation* of the *outcome* of interest.

Probabilistic sampling. A sampling process that assigns a well-defined probability for each *unit* to be drawn from a *sampling frame*. They include *random sampling*, *stratified random sampling*, and *cluster sampling*.

Process evaluation. An evaluation that focuses on how a program is implemented and operates, assessing whether it conforms to its original design and documenting its development and operation. Contrast with *impact evaluation*.

Propensity score. Within the context of *impact evaluations* using *matching* methods, the propensity score is the probability that a *unit* will enroll in the program based on observed characteristics. This score is a real number between 0 and 1 that summarizes the influence of all of the observed characteristics on the likelihood of enrolling in the program.

Propensity score matching. A *matching* method that relies on the *propensity score* to find a *comparison group* for a given *treatment group*.

Prospective evaluation. Evaluations designed and put in place before a program is implemented. Prospective evaluations are embedded into program implementation plans. Contrast with *retrospective evaluation*.

Quasi-experimental method. *Impact evaluation* methods that do not rely on *randomized assignment* of treatment. *Difference-in-differences, regression discontinuity design*, and *matching* are examples of quasi-experimental methods.

Randomized assignment or randomized controlled trials. *Impact evaluation* method whereby every eligible *unit* (for example, an individual, household, business, school, hospital, or community) has a probability of being selected for treatment by a program. With a sufficiently large number of *units*, the process of randomized assignment ensures equivalence in both observed and unobserved characteristics between the *treatment group* and the *comparison group*, thereby ruling out *selection bias*. Randomized assignment is considered the most robust method for estimating *counterfactuals* and is often referred to as the gold standard of *impact evaluation*.

Randomized promotion. *Instrumental variable* method to estimate program impacts. The method randomly assigns to a subgroup of units a *promotion*, or encouragement to participate in the program. Randomized promotion seeks to increase the take-up of a voluntary program in a randomly selected subsample of the population. The promotion can take the form of an additional incentive, stimulus, or information that motivates units to enroll in the program, without directly affecting the outcome of interest. In this way, the program can be left open to all eligible units.

Random sample. A sample drawn based on *probabilistic sampling*, whereby each unit in the *sampling frame* has a known probability of being drawn. Selecting a random sample is the best way to avoid an unrepresentative *sample*. Random sampling should not be confused with *randomized assignment*.

Regression analysis. Statistical method to analyze the relationships between a *dependent variable* (the variable to be explained) and *explanatory variables*. Regression analysis is not generally sufficient to capture causal effects. In *impact evaluation*, regression analysis is a way to represent the relationship between the value of an *outcome* indicator Y (dependent variable) and an independent variable that captures the assignment to the treatment or comparison group, while holding constant other characteristics. Both the assignment to the treatment and comparison group and the other characteristics are explanatory variables. Regression analysis can be univariate (if there is only one explanatory variable; in the case of impact evaluation, the only explanatory variable is the assignment to the treatment or comparison group) or multivariate (if there are several explanatory variables).

Regression discontinuity design (RDD). A *quasi-experimental* impact evaluation method that can be used for programs that rely on a continuous index to rank potential participants and that have a cutoff point along the index that determines whether potential participants are eligible to receive the program or not. The cutoff threshold for program eligibility provides a dividing point between the *treatment group* and the *comparison group*. Outcomes for participants on one side of the cutoff are compared with outcomes for nonparticipants on the other side of the cutoff. When all units comply with the assignment that corresponds to them on the basis of their eligibility index, the RDD is said to be "sharp." If there is noncompliance on either side of the cutoff, the RDD is said to be "fuzzy."

Results chain. Sets out the program logic by explaining how the development objective is to be achieved. It articulates the sequence of *inputs, activities,* and *outputs* that are expected to improve *outcomes*.

Retrospective evaluation. An evaluation designed after a program has been implemented (ex post). Contrast with *prospective evaluation.*

Sample. In statistics, a sample is a subset of a *population of interest*. Typically, the population is very large, making a *census* or a complete enumeration of all the values in the population impractical or impossible. Instead, researchers can select a representative subset of the population (using a *sampling frame*) and collect statistics on the sample; these may be used to make inferences or to extrapolate to the population. This process is referred to as *sampling.* Contrast with *census.*

Sampling. A process by which units are drawn from a *sampling frame* built from the *population of interest*. Various alternative sampling procedures can be used. *Probabilistic sampling* methods are the most rigorous because they assign a well-defined probability for each unit to be drawn. *Random sampling*, *stratified random sampling*, and *cluster sampling* are all probabilistic sampling methods. Nonprobabilistic sampling (such as purposive or convenience sampling) can create sampling errors.

Sampling frame. A comprehensive list of units in the *population of interest*. An adequate sampling frame is required to ensure that the conclusions reached from analyzing a sample can be generalized to the entire population. Differences between the sampling frame and the population of interest create a *coverage bias*. In the presence of coverage bias, results from the sample do not have *external validity* for the entire population of interest.

Selection. Occurs when program participation is based on the preferences, decisions, or unobserved characteristics of participants or program administrators.

Selection bias. The estimated *impact* suffers from selection bias when it deviates from the true *impact* in the presence of *selection.* Selection bias commonly occurs when unobserved reasons for program participation are correlated with *outcomes*. This bias commonly occurs when the *comparison group* is ineligible or self-selects out of treatment.

Sensitivity analysis. How sensitive the analysis is to changes in the assumptions. In the context of *power calculations*, it helps statisticians to understand how much the required *sample* size will have to increase under more conservative assumptions (such as lower expected impact, higher variance in the outcome indicator, or a higher level of *power*).

Significance. Statistical significance indicates the likelihood of committing a *type I error*, that is, the likelihood of detecting an impact that does not actually exist. The significance level is usually denoted by the Greek symbol α (alpha). Popular levels of significance are 10 percent, 5 percent, and 1 percent. The smaller the significance level, the more confident you can be that the estimated impact is real. For example, if you set the significance level at 5 percent, you can

be 95 percent confident in concluding that the program has had an impact if you do find a significant impact.

Significance test. A test of whether the *alternative hypothesis* achieves the predetermined *significance* level in order to be accepted in preference to the *null hypothesis*. If a test of significance gives a *p* value lower than the statistical significance (α) level, the null hypothesis is rejected.

SMART: Specific, measurable, attributable, realistic, and targeted. Good *indicators* have these characteristics.

Spillovers. Occur when the treatment group directly or indirectly affects outcomes in the comparison group (or vice versa).

Stable unit treatment value assumption (SUTVA). The basic requirement that the *outcome* of one *unit* should be unaffected by the particular assignment of treatments to other units. This is necessary to ensure that *randomized assignment* yields unbiased estimates of *impact*.

Statistical power. The *power* of a statistical test is the probability that the test will reject the *null hypothesis* when the *alternative hypothesis* is true (that is, that it will not make a *type II error*). As power increases, the chances of a type II error decrease. The probability of a type II error is referred to as the false negative rate (β). Therefore power is equal to $1 - \beta$.

Stratified sample. Obtained by dividing the population of interest (*sampling frame*) into groups (for example, male and female), and then drawing a *random sample* within each group. A stratified sample is a probabilistic sample: every *unit* in each group (or stratum) has a known probability of being drawn. Provided that each group is large enough, stratified sampling makes it possible to draw inferences about outcomes not only at the level of the population but also within each group.

Substitution bias. An unintended behavioral effect that affects the *comparison group*. *Units* that were not selected to receive the program may be able to find good substitutes for the treatment through their own initiative.

Survey data. Data that cover a *sample* of the population of interest. Contrast with *census* data.

Synthetic control method. A specific matching method that allows statisticians to estimate impact in settings where a single *unit* (such as a country, a firm, or a hospital) receives an *intervention* or is exposed to an event. Instead of comparing this treated unit to a group of untreated units, the method uses information about the characteristics of the treated unit and the untreated units to construct a synthetic, or artificial, comparison unit by weighing each untreated unit in such a way that the synthetic comparison unit most closely resembles the treated unit. This requires a long series of observations over time of the characteristics of both the treated unit and the untreated units. This combination of comparison units into a synthetic unit provides a better comparison for the treated unit than any untreated unit individually.

Theory of change. Explains the channels through which programs can influence final *outcomes*. It describes the causal logic of how and why a particular program, program modality, or design innovation will reach its intended outcomes. A theory of change is a key underpinning of any *impact evaluation*, given the cause-and-effect focus of the research.

Time-invariant factor. Factor that does not vary over time; it is constant.

Time-varying factor. Factor that varies over time.

Treatment. See *intervention.*

Treatment group. Also known as the *treated group* or the *intervention group*. The treatment group is the group of *units* that receives an *intervention*, versus the *comparison group* that does not.

Treatment-on-the-treated (TOT). TOT estimates measure the difference in outcomes between the units that actually receive the treatment and the comparison group.

Type I error. Also known as a *false positive* error. Error committed when rejecting a *null hypothesis,* even though the null hypothesis actually holds. In the context of an *impact evaluation*, a type I error is made when an evaluation concludes that a program has had an *impact* (that is, the null hypothesis of no impact is rejected), even though in reality the program had no impact (that is, the null hypothesis holds). The *significance level* is the probability of committing a type I error.

Type II error. Also known as a *false negative* error. Error committed when accepting (not rejecting) the *null hypothesis,* even though the null hypothesis does not hold. In the context of an *impact evaluation*, a type II error is made when concluding that a program has no *impact* (that is, the null hypothesis of no impact is not rejected) even though the program did have an impact (that is, the null hypothesis does not hold). The probability of committing a type II error is 1 minus the *power* level.

Unit. A person, a household, a community, a business, a school, a hospital, or other unit of observation that may receive or be affected by a program.

Unit nonresponse. Arises when no information is available for some subset of units, that is, when the actual sample is different than the planned sample.

Unobserved variables. Characteristics that are not observed. These may include characteristics such as motivation, preferences, or other personality traits that are difficult to measure.

Variable. In statistical terminology, a symbol that stands for a value that may vary.

ECO-AUDIT

Environmental Benefits Statement

The World Bank Group is committed to reducing its environmental footprint. In support of this commitment, the Publishing and Knowledge Division leverages electronic publishing options and print-on-demand technology, which is located in regional hubs worldwide. Together, these initiatives enable print runs to be lowered and shipping distances decreased, resulting in reduced paper consumption, chemical use, greenhouse gas emissions, and waste.

The Publishing and Knowledge Division follows the recommended standards for paper use set by the Green Press Initiative. The majority of our books are printed on Forest Stewardship Council (FSC)–certified paper, with nearly all containing 50–100 percent recycled content. The recycled fiber in our book paper is either unbleached or bleached using totally chlorine-free (TCF), processed chlorine-free (PCF), or enhanced elemental chlorine-free (EECF) processes.

More information about the Bank's environmental philosophy can be found at HYPERLINK http://www.worldbank.org/corporateresponsibility http://www.worldbank.org/corporateresponsibility.

"*Impact Evaluation in Practice* is simply a gem. It encourages an approach to impact evaluation that seeks to be scientifically credible while at the same time recognizing the practical realities of doing this kind of work on the ground. There are valuable insights along these two dimensions throughout the entire book. I assign readings from this book all the time when training professionals interested in conducting, commissioning, or consuming impact evaluations."

—**Dan Levy,** *Senior Lecturer in Public Policy and Faculty Chair of the Strengthening Learning and Teaching Excellence Initiative, Kennedy School of Government, Harvard University*

"*Impact Evaluation in Practice* is a major contribution to the contemporary development agenda. It is an extremely valuable resource for evaluators in governments and development agencies, as well as universities and think tanks."

—**Leonard Wantchekon,** *Professor of Politics and International Affairs, Princeton University; Founder and President of the African School of Economics*

"The aim of this book is to provide an accessible, comprehensive, and clear guide to impact evaluation. The material, ranging from motivating impact evaluation, to the advantages of different methodologies, to power calculations and costs, is explained very clearly, and the coverage is impressive. This book will become a much consulted and used guide and will affect policy making for years to come."

—**Orazio Attanasio,** *Professor of Economics, University College of London; Director, Centre for the Evaluation of Development Policies, Institute of Fiscal Studies, United Kingdom*

"The updated version of this extraordinary book comes at a critical time—the culture and interest in evaluation is growing and needs to be supported with good technical work. *Impact Evaluation in Practice* is an essential resource for evaluators, social programs, ministries, and others committed to making decisions using good evidence. This work is increasingly important as the global development community works to reduce poverty and achieve the 2030 Sustainable Development Goals."

—**Gonzalo Hernandez,** *Executive Secretary, National Council for the Evaluation of Social Development Policy, Mexico*

Additional information is available on the *Impact Evaluation in Practice* website at http://www.worldbank.org/ieinpractice.

SIEF Strategic Impact Evaluation Fund

IDB Inter-American Development Bank

WORLD BANK GROUP

ISBN 978-1-4648-0779-4

SKU 210779