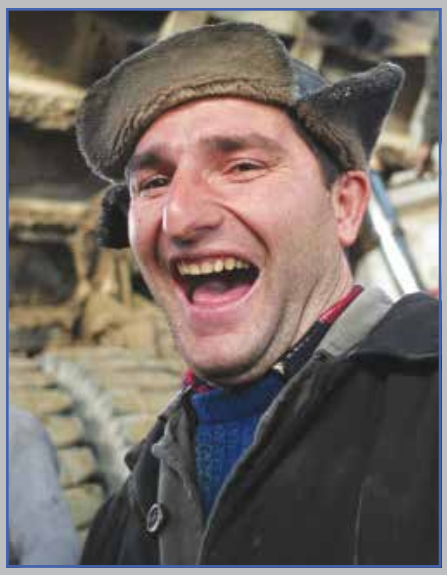
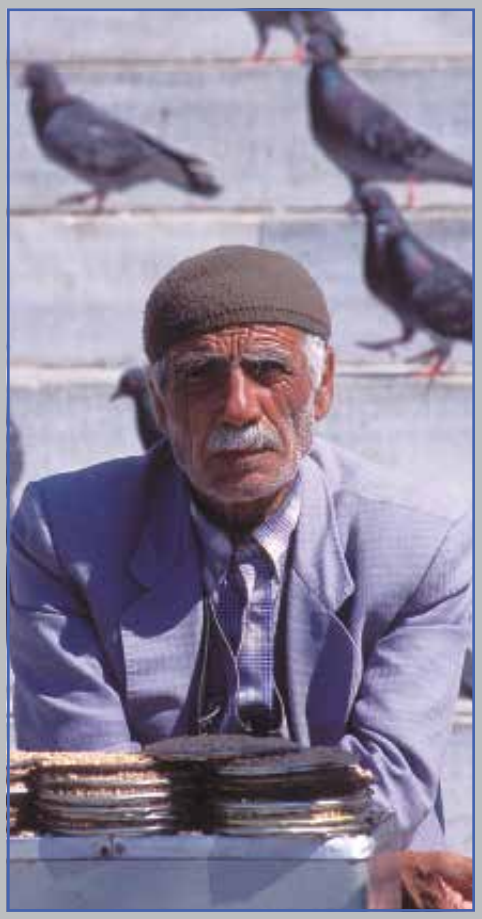
79808
June 2013

FINANCIAL LITERACY & EDUCATION
RUSSIA TRUST FUND





A **TOOLKIT**
for the **evaluation** of financial capability programs in low- and middle-income countries

A toolkit for the evaluation of financial capability programs in low- and middle-income countries

JOANNE YOONG
KATA MIHALY,
SEBASTIAN BAUHOFF
LILA RABINOVICH
ANGELA HUNG
of THE RAND CORPORATION

with support from ELAINE KEMPSON

FINANCIAL LITERACY & EDUCATION
RUSSIA TRUST FUND

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

Cover photos: World Bank
Cover design/layout: Nita Congress

Contents

Preface — xi

Acknowledgments — xii

1 Financial capability: concepts and approaches — 1
- 1.1 Defining financial capability — 3
- 1.2 Financial capability programs: what do we know? — 7
- 1.3 The Russia Financial Literacy and Education Trust Fund: new evidence and lessons for practice — 10
- 1.4 Overview of the Toolkit — 13

Further reading — 14

PART I SETTING THE STAGE FOR M&E: UNDERSTANDING THE M&E PROCESS AND CONCEPTS

2 Monitoring and evaluation — 21
- 2.1 What is monitoring? — 21
- 2.2 What is evaluation? — 22
  - 2.2.1 Process evaluation — 23
  - 2.2.2 Impact evaluation — 24
  - 2.2.3 Cost analyses — 26
- 2.3 Deciding when and how to undertake M&E — 27
- 2.4 Making the decision to evaluate — 28
  - 2.4.1 What type of evaluation is desirable? — 29
  - 2.4.2 What type of evaluation is possible? — 30
  - 2.4.3 An evaluability assessment — 30

Key points — 31

Further reading — 32

3 Setting the stage — 33
- 3.1 How should evaluators go about formalizing program goals and objectives? — 33
- 3.2 How can evaluators capture a program's theory of change? — 36
- 3.3 How can evaluators develop effective indicators? — 39
- 3.4 Using existing resources for indicators — 42

3.4.1 Measuring financial capability: the World Bank financial capability survey — 43
3.4.2 Measuring financial inclusion: the Global Findex questionnaire — 43
Key points — 44
Further reading — 44

PART II CONDUCTING M&E FOR FINANCIAL CAPABILITY PROGRAMS

4 Monitoring — 47
4.1 What is monitoring and why should we do it? — 47
4.2 How can we develop and implement a monitoring plan? — 48
4.3 How do we bring monitoring information together? — 52
4.4 How do we use monitoring in project management? — 55
4.5 How can we build a robust monitoring system? — 56
Key points — 57
Further reading — 57
General — 57
Technical — 57

5 Process evaluation — 59
5.1 When should we conduct a process evaluation? — 59
5.2 How should process evaluations be conducted? — 60
5.2.1 Develop results framework — 61
5.2.2 Develop process evaluation questions — 61
5.2.3 Collect necessary data — 61
5.2.4 Analyze data — 65
5.2.5 Develop and implement solutions to problems and challenges — 65
Key points — 65
Further reading — 66
General — 66
Technical — 66

6 Impact evaluation — 67
6.1 Key concepts for impact evaluation — 68
6.2 A caution against false methods — 71
6.3 Experimental evaluation designs — 73
6.3.1 Randomized control trial — 73
6.3.2 Encouragement design — 79
6.4 Quasi-experimental evaluation designs — 83
6.4.1 Regression discontinuity design — 83
6.4.2 Propensity score matching — 86
6.4.3 Difference-in-difference — 91
6.5 Combining impact evaluation methods — 95

6.6 Practical and logistical challenges to impact evaluation — 96
6.6.1 Compromises implementing and maintaining randomization — 96
6.6.2 Difficulty with eligibility rules and protocols — 97
6.6.3 Maintaining control group integrity — 97
Key points — 98
Further reading — 99
General — 99
Technical — 99

7 Putting it all together — 101
7.1 Why is a comprehensive evaluation so important? — 101
7.2 How do we conduct comprehensive evaluations? — 102
7.3 When is a comprehensive approach most critical? — 103
7.3.1 Conducting the exploratory (or pilot) phase of an evaluation — 104
7.3.2 Following up after impact evaluation and understanding differences in program effects — 105
7.3.3 Increasing the validity of evaluation findings — 105
7.3.4 Case studies of small or very sensitive populations — 106
Key points — 108
Further reading — 108
General — 108
Technical — 109

PART III COLLECTING AND ANALYZING M&E DATA FOR FINANCIAL CAPABILITY PROGRAMS

8 Data collection methods — 113
8.1 Qualitative data collection methods — 113
8.1.1 In-depth interviews — 114
8.1.2 Semi-structured and cognitive interviews — 115
8.1.3 Focus groups — 118
8.1.4 Desk review of documents and materials — 122
8.1.5 Audit or mystery shopping studies — 124
8.2 Quantitative data collection methods — 125
8.2.1 Surveys — 126
8.2.2 Site visits and observation — 129
8.2.3 Using existing administrative data — 131
8.3 Comparing the uses of different data collection methods — 133
Key points — 133
Further reading — 135
General — 135
Technical — 135

9 The process of collecting data: practical guidance — 137
9.1 Thinking about the process of data collection — 137
9.2 Sampling design and selection — 139
9.2.1 Sampling for quantitative research — 139
9.2.2 Sampling for qualitative research — 143
9.3 Choosing the mode of data collection — 146
9.4 Pilot testing — 147
9.5 Field implementation — 148
9.6 Data management, access, and documentation — 153
9.7 Gaining access to sensitive financial information — 154
Key points — 155
Further reading — 155
General — 155
Technical — 156

10 Analyzing quantitative and qualitative data — 157
10.1 Analyzing quantitative data — 157
10.1.1 Descriptive statistics — 157
10.1.2 inferential statistics and hypothesis testing — 163
10.1.3 Regression analysis — 165
10.1.4 Estimating and interpreting program effects in practice — 167
10. 2 Analyzing qualitative data — 172
10.2.1 Review the data and develop descriptive themes — 173
10.2.2 Code and formalize the data — 174
10.2.3 Conduct systematic analysis — 174
Key points — 176
Further reading — 176
General — 176
Technical — 176

PART IV OTHER ISSUES IN CONDUCTING M&E FOR FINANCIAL CAPABILITY PROGRAMS

11 Cost analysis: weighing program costs and benefits — 179
11.1 What are the steps in conducting a cost analysis? — 180
11.2 Specifying program alternatives — 180
11.3 Choosing a perspective (or several) — 181
11.4 Measuring benefits and costs — 181
11.5 Comparing alternatives — 183
11.5.1 Cost-benefit analysis — 183
11.5.2 Cost-effectiveness analysis — 184
11.5.3 Cost-consequences analysis — 186
Key points — 187

Further reading — 188
General — 188
Technical — 188

12 Implementing the evaluation — 189
12.1 Logistics and timing — 189
12.2 Forming an evaluation team — 190
12.3 Budget and financing — 194
12.4 Estimating total evaluation costs — 195
12.5 Exploring funding sources — 196
12.6 Contingency planning — 196
12.7 Preparing and reviewing a formalized evaluation plan — 197
Key points — 199
Further reading — 199
General — 199
Technical — 199

13 Ethical considerations — 201
13.1 Ethical Issues in a financial capability program setting — 201
13.2 Informed consent — 203
13.3 Confidentiality — 205
13.4 Anonymity — 206
13.5 Risk assessment and mitigation planning — 207
13.6 Ethics in evaluations with randomized interventions — 208
13.7 Other ethical obligations — 210
Key points — 211
Further reading — 212
General — 212
Technical — 212

14 Documenting and communicating results — 213
14.1 The importance of documentation — 214
14.2 Thinking about the relevant audiences — 214
14.2.1 Internal audiences — 215
14.2.2 External audiences — 216
14.3 Communicating with key audiences — 217
14.3.1 Summary — 218
14.3.2 Introduction — 220
14.3.3 Program description — 220
14.3.4 Evaluation design and methods — 220
14.3.5 Findings/results — 221
14.3.6 Conclusions and recommendations — 222
14.4 Dissemination: getting the word out — 223
14.5 Putting it all together — 226
Key points — 226

Further reading — 226
General — 226
Technical — 226

Appendixes — 229
A Trust Fund financial capability evaluation projects — 231
B Technical appendix — 247

Boxes

4.1 Monitoring in practice: an example from the RTF pilot program in Malawi — 52
5.1 RTF in practice: Process Evaluation using surveys — 63
5.2 RTF in practice: process evaluation using focus groups and interviews — 64
6.1 Selection bias — 70
6.2 What is the role of qualitative research in impact evaluation? — 72
6.3 RTF in practice: designing randomization to avoid spillovers — 75
6.4 RTF in practice: example of an RCT in Brazil — 76
6.5 Randomized phase-in — 78
6.6 Instrumental variables — 80
6.7 RTF in practice: example of encouragement design, South Africa — 81
6.8 A note on treatment effects — 82
6.9 Example of RDD in practice — 83
6.10 Example of PSM in practice — 90
6.11 RTF in practice: applying DID to the evaluation of a savings program, Nigeria — 92
6.12 RTF in practice: school-based financial education in Brazil — 95
6.13 RTF in practice: example of the Hawthorne effect — 98
7.1 RTF in practice: comprehensive evaluation of financial training program in India — 104
8.1 Topic guide extracts for in-depth interviews — 116
8.2 Using cognitive interviewing in the developmental phase of an RTF program — 117
8.3 Using focus groups in RTF programs — 120
8.4 Desk review to assess financial advice — 123
8.5 Audit studies in practice: an example from an RTF pilot program in Mexico — 124
8.6 Household survey design in developing countries — 127
8.7 Using administrative data in practice: RTF projects — 131
8.8 Administrative data for evaluating financial capability programs — 132
9.1 Challenges of obtaining a sample frame — 140
9.2 Convenience sampling — 144
10.1 Hypothesis testing: example from the RTF pilot program on school-based financial education in Brazil — 164
10.2 Example of partial compliance — 169
11.1 Financial education versus subsidies and money-back guarantees in Indonesia and India — 185

12.1 The opportunities of working with a hired survey company: example from an RTF pilot — 193
12.2 The challenges of working with hired survey companies: an example from Indonesia — 194
13.1 Dealing with disappointment in control groups — 209
13.2 RTF in practice: exploiting phased roll-outs in Nigeria — 210
14.1 Reporting results in journal articles — 223

Figures

1.1 A conceptual model of financial capability — 6
3.1 Examples of goals versus objectives — 36
3.2 An example of a results framework model — 37
4.1 Example indicator protocol reference sheet — 51
4.2 Pyramid of monitoring information flows — 53
5.1 Steps in conducting a process evaluation — 60
5.2 Sample questions for process evaluation — 62
6.1 Decision tree for choosing impact evaluation design — 71
6.2 Three steps in randomized assignment — 74
6.3 Propensity scores of treatment and comparison groups — 89
6.4 Graphical illustration of the DID design — 92
8.1 Kinds of documents to review — 123
9.1 The process of data collection — 138
9.2 Survey types — 139
10.1 Histogram of land holdings in Village A — 160
10.2 Scatter plot of household savings and income — 162
10.3 Hypothesis testing — 163
10.4 A rigorous approach to qualitative analysis — 173
11.1 Basic steps for cost analysis — 180
11.2 Key cost categories — 183
12.1 Measuring performance and effectiveness: example of linking to the program cycle — 191
14.1 Documentation and communications flow chart — 213

Tables

1.1 RTF-funded pilot projects — 11
1.2 A reader's guide to the Toolkit — 15
2.1 Sample monitoring checklist — 22
2.2 Monitoring, process evaluation, and impact evaluation — 31
3.1 Expected outcomes, activities, and indicators in the *Scandal!* program — 40
3.2 Expected outcomes, activities, and indicators in the feature film program in Nigeria — 41
4.1 Example monitoring implementation plan — 49
4.2 Checklist for ensuring a quality monitoring system — 56
6.1 Implementing difference-in-differences — 94
6.2 Overview of impact evaluation methods — 99

7.1 Checklist for deciding on a mixed-methods evaluation design — 108
8.1 Sample interview question set for program beneficiaries and key informants — 118
8.2 Checklist for survey instrument design — 129
8.3 Comparing key qualitative and quantitative data collection methods — 134
9.1 Benefits and limitations of sampling methods — 142
10.1 Contingency — 161
10.2 Contingency with continuous variable — 161
11.1 Example of a CCA that compares alternative financial capability programs — 186
11.2 Types of cost analyses — 187
12.1 Core evaluation team roles and responsibilities/skills — 192
12.2 Data collection evaluation team roles and responsibilities — 192
12.3 Example budget from the Nigerian RTF pilot — 195
12.4 Evaluation plan checklist — 198
13.1 Elements of data-safeguarding plans — 206
14.1 Generic Template for a Formal Report — 219
14.2 Sample dissemination plan — 225
14.3 Key issues in communicating and disseminating evaluation results — 227

Preface

When resources are scarce and social safety nets are weak, households' ability to manage income and assets wisely may be an important determinant of economic security. However, many open questions remain about how households in low- and middle-income countries gain and exercise financial capability, and the best ways for governments and the private/nonprofit sector to help increase this capability. With the exception of a small but important number of studies that have recently been completed or are currently under way, robust evidence regarding the efficacy of financial capability interventions is relatively sparse compared to the level of interest and programmatic activity. One reason for this is a lack of systematic evaluation.

While there are many useful toolkits that address different aspects of program evaluation, this Toolkit, sponsored by the World Bank's Russia Financial Literacy and Education Trust Fund, focuses on the specific challenges of evaluating financial capability interventions in the low-income and middle-income country setting. A comprehensive approach to high-quality evaluation seeks to identify effects that can be causally attributed to a program, to draw conclusions about these effects that reflect both program theory and the realities of implementation, and to derive insights that can be reasonably generalized to a broader population. This Toolkit provides a practical guide to the various aspects of conducting such evaluations that addresses the specific needs of financial capability programs in the developing world.

The Toolkit draws from past experience and the experience of the Russia Financial Literacy and Education Trust Fund pilot projects to provide concrete and tangible examples for the reader that illustrate the specific circumstances and challenges in this field. For further reference, the reader is invited to adapt the templates and survey instruments provided freely. Our hope is that the material presented here supports the efforts of a diverse group of users, and contributes to the generation of an ever more robust body of evidence on how to best support financial consumers around the developing world in their quest to achieve financial security.

Acknowledgments

This publication is the result of an extensive cooperative effort over the course of three years under the aegis of the World Bank's Russia Financial Literacy and Education Trust Fund. In addition to authors, the project involved the intellectual contributions and efforts of many other individuals without whom it would not have been possible. The project was overseen by the World Bank's project team which included Mattias Lundberg as Task Team Leader supported by Florentina Mulaj, who managed the coordinated program of evaluation studies that were integral to the effort. Robert Holzmann and Richard Hinz, who have collaborated over the years in managing the Trust Fund for the World Bank, provided overall guidance, feedback, and support. Arie Kapteyn directed the work of the Rand Corporation staff.

A wide range of others provided advice, content, examples, and suggestions that have contributed to the publication. These include Annamaria Lusardi (George Washington University), Robert Walker (Oxford University), Gerrit Antonides (Wageningen University), Sharon Collard (Bristol University), Polly Jones, Kinnon Scott, Leora Klapper, Xavi Giné, Gunhild Berg, Bilal Zia, Giuseppe Iarossi, Rogelio Marchetti, Aidan Coville, David McKenzie, Leopold Sarr, Valeria Perotti, Margaret Miller, and Nathan Fiala (all of whom are World Bank staff or consultants). We are also grateful to our colleagues from the OECD—Andre Laboul, Bruno Levesque, and Flore-Anne Messy—for their support and feedback; and to colleagues from DfID's Financial Education Fund, in particular Alyna Wyatt, for their generous sharing of ideas and content from the related effort.

We are also extremely grateful to our peer reviewers Audrey Pettifor (University of North Carolina), Rajiv Prabhakar (London School of Economics), Billy Jack (Georgetown University), Lew Mandell (University of Washington), Christel Vermeersh (World Bank), Miriam Bruhn (World Bank), and Patrick Premand (World Bank), for their guidance and challenging suggestions. Raiden Dillard of the World Bank and Nita Congress managed the layout and publication of this document, and Amira Nikolas provided invaluable support for the myriad administrative tasks required for the effort.

Finally, we thank the participants in the many workshops and conferences sponsored by the Trust Fund held in Vienna, Paris, Montevideo, Washington, Cape Town, St. Petersburg, Cartegena, Nairobi, and New Delhi, whose questions and comments as the work developed greatly enhanced the organization and content of this guide.

CHAPTER 1

Financial capability: concepts and approaches

Over the last 10–15 years, there has been an increasing awareness that the ability of individuals to achieve personal financial well-being can have serious implications, both for those individuals and for society as a whole. A range of interventions in a series of policy areas can impact the ability of individuals to achieve personal financial well-being. These policy areas include financial inclusion, consumer protection, over-indebtedness, and social protection (particularly in areas where consumers have to make their own provision for it). And financial capability programs in these policy areas can be useful tools in helping to achieve this impact.

In 2008, the Russian Federation and the World Bank signed an agreement to establish a Trust Fund at the World Bank to support the advancement of financial capability programs. The Russia Financial Literacy and Education Trust Fund (RTF) objectives are to develop a comprehensive definition of financial capabilities, review current research and information on existing programs, and use this to develop and test measurement and evaluation methods. To achieve these objectives, the RTF sponsored a large program of work, including this Toolkit. Other activities include conducting a research initiative to define and measure financial capability and supporting evaluations of innovative financial capability programs in low- and middle-income countries. Meeting these objectives is expected to lead to standardized methods to measure knowledge and capabilities, design and operate financial capability programs, and evaluate their effectiveness in helping individuals effectively manage their interactions with a wide range of financial services and products. The knowledge derived from this effort will be widely disseminated to enhance the capacity of countries to develop effective financial capability and education strategies.

Interest in this area has evolved in two important ways. First, supported by a growing body of research, the area has expanded from a relatively narrow focus on financial literacy or knowledge to a broader range of interests relating to consumer behavior. A new term—**financial capability**—was coined to reflect this change in emphasis and as a shorthand term that captures a range of different competencies.

Second, while interest in improving individuals' financial capabilities was initially concentrated in high-income countries, this interest has expanded to poorer parts of

the world, including both low- and middle-income countries (LMIC). Although people living on low incomes in a low-income country can have very sophisticated financial lives that may not necessarily require interaction with financial services, concerns can be subtly different in such settings. Research carried out as part of the RTF confirms this conclusion, while reaffirming the existence of core competencies that are applicable across income levels and across countries.

And there has been a shift in the types of intervention used to address financial capability, mirroring the shift in focus from increasing knowledge to changing behaviors. Early programs tended to be education-based and, typically, used workshops with adults and school-based teaching for children and young people. A shift in focus recognizes that education alone (in its narrowest sense) may not be the best (or even the most appropriate) way of influencing behaviors. This has translated more recently into a wider range of interventions that seek to modify behaviors rather than just inculcate information. These range from one-on-one guidance or counseling to social marketing, edutainment, and the use of specific design elements in general financial information and disclosure materials for consumers.

Unfortunately, despite this burgeoning interest and rapid growth in the number of financial capability interventions of all kinds, there is a relatively small evidence base on their relative efficacy to help support developments in policy and practice, be it a national strategy to tackle low levels of financial capability or the design of a specific intervention at a local level. In fact, few interventions have been evaluated at all, and where they have been, the design of the evaluations has tended to limit what others can learn from them. Significantly, very few evaluations rigorously measure the impact of an intervention and even fewer explain how and why the impact (or lack of it) discovered during the intervention occurred.

This Toolkit is designed for researchers who are interested in conducting an evaluation of a financial capability program and for policy makers and practitioners interested in commissioning an evaluation. It will also be useful to evaluation researchers who want to brush up on a research technique they are less familiar with or who are new to the area of financial capability and financial education, particularly in LMIC. The Toolkit does not provide guidance on which financial capability programs are effective or appropriate. Instead, it aims to provide tools for carefully and rigorously monitoring and evaluating financial capability programs, as well as recommendations on choosing appropriate methods and their proper application.

This chapter begins by setting out what we mean when we refer to “financial capability” and “financial education”; it reviews the existing evidence on the effectiveness and impact of programs to raise levels of financial capability and provides an overview of interventions that have been supported by the RTF. The chapter concludes with a discussion of how the Toolkit is organized and how it can be used.

1.1 DEFINING FINANCIAL CAPABILITY

The Organisation for Economic Co-operation and Development (OECD) conducted a review that found that although there is little consensus about whether to use the original term "financial literacy" or the more recently coined "financial capability," the definitions of these two terms differ little; the review also found that many draw on the definition of financial literacy that was first put forward in 1992 by Noctor et al. (1992). Financial literacy is defined as the ability to make informed judgments and to take effective decisions regarding the use and management of money.

Although the term financial literacy was originally used in a fairly narrow sense to refer primarily to financial knowledge and understanding, this quickly became too limiting and was broadened over time to include financial skills and competencies, attitudes, and behavior. In Italy and Spanish-speaking countries the term "financial culture" was adopted; the Netherlands has used the term "financial insight," while "financial competencies" is the preferred term in the Pacific Islands. The U.K. Regulator (then the Financial Services Authority) adopted the term "financial capability" to reflect this broadened concept and to distance the term from basic skills (numeracy and literacy), because research had shown that even well-educated people with high incomes can be financially incapable. Over time, the U.K. term has tended to be the one that is used most widely.

The terminology used is one thing, but more fundamental is a divide in how this important area of interest is conceptualized, a divide that manifests itself in the design and content of the interventions themselves and in their evaluation. It is also apparent in the design of surveys to measure levels of financial capability or literacy of a population. Holzmann, Mulaj, and Perotti (2013) identify two "polar views": the **cognitive/normative approach** and the **outcome-oriented/positive approach**.

The cognitive/normative approach assumes that a lack of financial knowledge is a key constraint and that appropriate knowledge and the ability to apply it determines appropriate behaviors (and outcomes). Although the approach acknowledges that attitudes play a part, it is implicit that education can improve behaviors by having an impact on knowledge, skills, and attitudes. Moreover, the knowledge, skills, and behavior that need to be influenced are determined normatively—that is, they are focused on what it is thought that people **should** know, believe, or do.

In contrast, the outcome-oriented/positive approach is more focused on behavior or outcomes and adopts an empirical approach to determining the behaviors that should be considered capable and those that should be considered less capable. This empirical rather than normative approach has been informed by research undertaken in the United Kingdom that showed that the normative approach to defining

financial capability did not reflect what ordinary people considered it to be. This is particularly important when considering financial capability in a low-income setting, where the norms of the people setting policy or delivering interventions to raise financial capability may not be in tune with the financial lives of the poor. Moreover, this approach accepts that behavior may be influenced by many things, including:

- **Personal factors**, such as knowledge and awareness and skills, but also trust, attitude, motivation, and behavioral biases
- **Societal factors** (or social norms)—what others do or believe
- **Environmental factors**, such as barriers to access to financial services, the behavior of financial firms, consumer protection legislation, and poverty.

Consequently, the outcome-oriented approach acknowledges a much wider range of policy interventions than education alone. Conventional financial education can address levels of knowledge or skills of individuals; social marketing and edutainment may, however, be more appropriate ways of changing trust, attitude, motivation, and even social norms.

Overcoming behavioral biases requires a rather different approach, such as designing interventions that harness rather than seek to overcome the prevalent biases. The most commonly cited example is using auto-enrollment into a pension fund to overcome the inertia many people exhibit in providing for their old age; the approach also harnesses inertia by requiring people to opt out if they do not want to be part of a pension scheme.

The RTF research initiative has adopted the outcome-oriented/positive approach. It began with over 70 focus groups held across eight LMICs[1] to ask people of all income levels to describe behaviors that constitute financial capability, in their own terms. Focus group results showed a remarkable level of consistency across countries:

- First and foremost, respondents reported that being financially capable means managing day-to-day spending well, including planning how money will be spent and sticking to it, knowing how much you have spent and how much money you have left, living within your means, and not spending money you do not have on non-essentials.
- Respondents also said that looking ahead and planning for major anticipated expenditures and providing for unexpected financial demands for one's children's futures and one's own old age was important.

[1] Papua New Guinea, Malawi, Zambia, Namibia, Tanzania, Uruguay, Colombia, and Mexico.

- And respondents said that choosing and using financial products wisely was also important, but only for those engaged in using financial services.

In other words, financial capability was described primarily in terms of behaviors. Surveys designed to capture these aspects of financial management and decision making show that while the term "financial capability" is a useful shorthand term, it encompasses a group of competencies that are not necessarily correlated with one another.

The focus groups also identified a range of factors that do and do not influence outcomes:

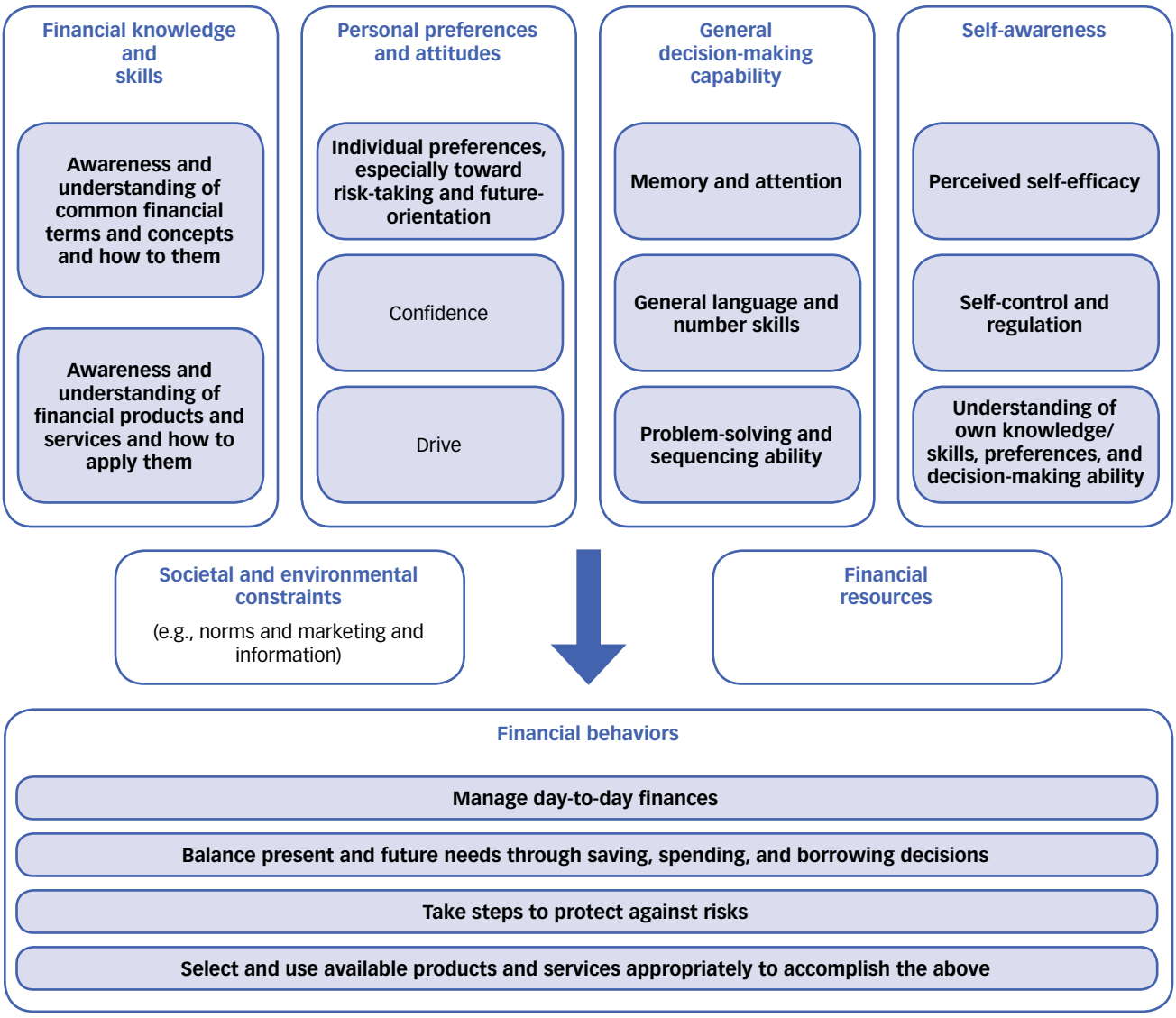
- Motivation and behavioral biases—such as having short-term horizons ("living for today"), impulsivity, wanting to get on in life, an ability to resist peer pressure or temptation, and self-control—play a very important part in determining a person's level of financial capability.
- There are capable and incapable people at all educational levels, and what is known at the time is less important than motivations and personality traits. In other words, financially capable people figure out what is best in their circumstances, regardless of their educational level or prior knowledge.
- People may not conform to the norm, but their behavior may still be considered capable. For example, many ways of providing for the future were described that had nothing to do with saving within a savings account, such as investing in livestock or buying building materials over an extended period of time to build a home at some time in the future.
- Very low income can restrict the choices people are able to make and may prevent them from demonstrating their true level of capability. The most commonly mentioned examples were living within one's means and planning for the long-term. But there are capable and incapable people at every income level.
- Social and environmental factors can also prevent people from demonstrating their true level of capability. Such factors include financial demands of family or other social networks, access to financial products, poor financial services provision, and inadequate consumer protection infrastructure.

These are all very important points when it comes to designing and evaluating interventions to raise levels of financial capability, and they have informed the work undertaken by the World Bank under the RTF, including the content of this Toolkit.

Drawing on these findings, we have defined financial capability as **the internal capacity to act in one's own best financial interests, given the socioenvironmental conditions**. In other words, a financially capable person is one who is able

to make and implement financial decisions that are considered appropriate for his or her peers and community, given their external circumstances. Figure 1.1 presents a conceptual model of financial capability. The top four boxes are personal factors that comprise financial capability. These personal factors, together with societal and environmental factors and financial resources jointly determine financial behaviors.

FIGURE 1.1 A CONCEPTUAL MODEL OF FINANCIAL CAPABILITY



1.2 FINANCIAL CAPABILITY PROGRAMS: WHAT DO WE KNOW?

Here, and in the rest of the Toolkit, we refer to a program as the subject of evaluation. A program may itself consist of a set of one or more related interventions delivered together or separately. The scope of the Toolkit covers programs that include interventions aimed at improving consumers' own financial capabilities and focus at least partly on behavioral outcomes.

These include "traditional" financial capability programs or financial education programs for adults and children, which typically involve classroom-based instruction in workshops or schools, and a focus on the acquisition of knowledge as the primary agent of behavior change. They also include innovative new models for disclosure, edutainment, and social marketing that use insights from behavioral finance to improve the communication of financial information and messaging. Given the definition provided above, we do not consider interventions that expand access to new products **alone** to be financial capability programs, but we do consider counseling and advice and interventions that encourage less-able individuals to seek out the help they need.

We actually know very little about what works in financial capability programs, not just in developing countries but also on a global scale. While there is a robust body of evidence that individual financial capability or literacy leads to better behavior and outcomes in both developed and developing countries, few studies are able to convincingly show the ability of financial capability programs to affect financial behavior, and even fewer convincingly document how.

Even in the developed country context, evidence on financial capability programs has been mixed. While there has been a proliferation in the United States and elsewhere of programs, tools, and materials seeking to educate and inform the public about saving, investment, and retirement objectives, relatively few of these offerings have been submitted to formal evaluation or assessment of what works. Evidence of correlation is often available, but only a handful of programs offer strong causal evidence that education leads to better financial capability and/or improves financial behavior. Financial education has been shown to effectively increase financial knowledge and to improve financial attitude, motivation, and behavioral intent (Lyons 2005; Lyons, Palmer, Jayaratne, and Scherpf 2006). At the same time, however, an active and growing research literature (see, for example, Bernheim, Garrett, and Maki 2001; Duflo and Saez 2003; Lyons, Rachlis, Staten, and Xiao 2006; Cole and Shastry 2009; and lately Collins and O'Rourke 2011) continues to find no conclusive evidence on whether financial education or counseling programs can in practice effectively bring about actual behavioral change.

The evidence on specific features of successful interventions—whether content, delivery, or timing—that work best is even more limited. Mandell (2009) presents evidence that school-based financial education programs for younger students (namely, pre–high school students) hold more promise than high school financial education programs, although high school financial education programs may not have a lasting impact on financial literacy scores (as documented for example in Mandell 2006). Collins and O'Rourke (2011) find that face-to-face counseling is no more effective than other modes. Lusardi and her colleagues (2010) investigate which strategies, interventions, and delivery mechanisms work best for financial education. They argue that workplace financial education is effective, especially when it occurs during "teachable moments," such as providing retirement education at the start of a career or at retirement. They also encourage the use of informal education—learning that occurs outside of a formal education setting—as the most prevalent method by which adults gain education; such informal education can occur through a number of channels, such as entertainment, casual conversations with friends or family, or seeking out information via an Internet search. However, these strategies remain largely unproven, as Collins and O'Rourke (2011) note in their review.

An important reason for the absence of good evidence on the various strategies is that the field of program evaluation in this area is still developing (Martin 2007). While some obstacles are the result of inherent conceptual difficulties, others reflect shortcomings in planning or design. Two leading research gaps identified by Schuchardt et al. (2009) include the lack of agreed-upon outcome metrics and substantial differences in the nature and quality of existing methodology. In the first instance, the majority of programs conduct evaluations based not on outcomes but rather on output measures, such as the number of participants enrolled or the number of programs provided (Lyons, Palmer, Jayaratne, and Scherpf 2006); few measure subjective satisfaction, knowledge, self-confidence/efficacy, and intentions; and even fewer measure behaviors. Yet, in practice, even when programs are very similar and relatively well defined, comparison is often hampered by the lack of standardized, common benchmark measures. In the second case, many early studies were carried out without sufficient consideration of their empirical validity, by failing to specify a control or comparison group or to adequately account for or even acknowledge selection bias (see Fox and Bartholomae 2008; Lusardi 2004; Lusardi and Mitchell 2007b, 2008; Lyons, Palmer, Jayaratne, and Scherpf 2006; Hogarth 2006; Collins and O'Rourke 2009, 2011; Lyons 2005; and Lyons, Palmer, Jayaratne, and Scherpf 2006). The environment in developing countries has posed even more challenges for evaluators: Often, important government or program administrative data are unavailable or unsuitable; target populations may be small, dispersed, or difficult to reach; and

program resources are generally limited, leading to evaluation studies that lack the statistical power to detect small changes in target population behavior (Zia 2011).

Driven in part by growing awareness of the need for rigorous evaluation among policy makers, practitioners, and researchers, however, an increasing number of evaluations explicitly aim to understand the causal relationship between interventions and financial knowledge, attitudes, and behavioral outcomes.

This includes a new and exciting wave of evidence from developing countries, including both experimental and nonexperimental studies.

In one of the first large-scale randomized experiments of financial education, Cole, Sampson, and Zia (2011) find that financial literacy training related to opening a bank account had no effect on the general population but does result in more account openings among the initially less-knowledgeable. A parallel literature on small-enterprise business training also offers important insights: Karlan and Valdivia (2010) find that a business education program improves recordkeeping, though not profits, among microfinance borrowers in Peru. Bruhn and Zia (2011) find significant impacts of a five-session business-training program on business investment in Bosnia and Herzegovina (although not on business survival), but, in contrast, they find that the program appears to work better for the ex ante more financially literate. Giné and Mansuri (2011) find that an eight-day business training program led to increased business knowledge, better business practices, and improvements for microfinance clients in Pakistan, but only for men.

These seminal studies are critical first steps in establishing rigorous proof-of-concept for various interventions. However, as with all experimental studies, it is important to contextualize these findings, especially when attempting to generalize more broadly. In some instances, the external or ecological validity of the findings may be limited in other settings and populations, especially when the outcome measures and interventions were highly locally specific—an observation that points to the importance of qualitative research in impact evaluations in helping explain why and how the impacts (or lack thereof) were observed.

In addition to these studies, a significant contribution to the field has been made by the Department for International Development Financial Education Fund (FEF). FEF was set up specifically to fund robust evaluations and build capacity, targeting organizations working in Sub-Saharan Africa and working in a participatory manner with practitioners to evaluate both existing programs and new initiatives. As a result, the funded evaluations include a wide array of interventions, from the inclusion of a debt-management storyline in *Makutano Junction*, a popular television series, to the launch of preferential youth banking accounts in Uganda. The experiences of the FEF programs also vary widely in terms of evaluation scope and approach (including

both experimental and nonexperimental studies), and many of the evaluation studies reflect real-world challenges that occur as organizations begin to develop evaluation capacity. Importantly, while not all the FEF studies are able or even attempt to rigorously measure impact by the standards of randomized controlled trials (RCTs), many undertake the important task of qualitatively and quantitatively exploring the process of operating new programs or establishing new programs so that others may learn from them. More information and learning materials from FEF can be found at: http://www.financialeducationfund.com.

Finally, research in this area is helping to develop the evidence base on the use of alternative strategies: While most previous interventions examine fairly labor-intensive and traditional adult education programs, two new studies explore innovative pedagogical methods with some success. Carpena, Cole, Shapiro, and Zia (2011) examine a novel video-based, five-week personal financial training program in India and find individuals who received financial literacy training are 5 percentage points more likely to know the concept of a household budget, 17 percentage points more likely to know the minimum requirements to open a bank account, and 20 percentage points more likely to understand unproductive loans. Drexler, Fischer, and Schoar (2010) compare a more traditional training course based on imparting conceptual understanding with a simpler, shorter, rule-of-thumb–based training for microentrepreneurs in the Dominican Republic and find that the use of simplified rules-of-thumb increases the likelihood of proper recordkeeping.

1.3 THE RUSSIA FINANCIAL LITERACY AND EDUCATION TRUST FUND: NEW EVIDENCE AND LESSONS FOR PRACTICE

As part of its initiative, the RTF is funding a set of individual pilot programs to test financial capability interventions. These pilot programs further expand the horizons of available evidence on financial capability programs. The pilot programs test new and exciting models of intervention in different settings and use a variety of mediums. Table 1.1 summarizes the projects that are described in greater detail in appendix A and on the RTF website, www.finlitedu.org. This Toolkit will reflect on the experiences of these programs, as well as other seminal studies, to draw lessons for evaluators.

The RTF pilot programs deliver financial capability programs through social marketing or edutainment programs based on large marketing campaigns in television, radio, and other media. They use drama, radio programming, television "soap opera" programs, specialized films, street theater performance, Internet-based media/CD- or DVD-ROM modules, and comic books/animation. This form of outreach may better capture individuals' attention. Such approaches can target audiences who may be

TABLE 1.1 RTF-FUNDED PILOT PROJECTS

PROJECT NAME	COUNTRY	DELIVERY CHANNEL
Social marketing or edutainment programs		
Harnessing Emotional Connections to Improve Financial Decisions: Evaluating the Impact of Financial Education in Mainstream Media in South Africa	South Africa	Television soap opera
The Impact of Financial Literacy through Feature Films: Evidence from a Randomized Experiment in Nigeria	Nigeria	Feature film
Measuring the Impact of Financial Literacy on Savings and Responsible Credit Card Use: Evidence from a Randomized Experiment in Mexico	Mexico	Radio soap opera and digital storytelling videos
The Impact of Cartoons and Comics on the Effectiveness of Financial Education: Evidence from a Randomized Experiment in Kenya	Kenya	Classroom instruction and comic strips
Learning by Doing? Using Savings Lotteries and Social Marketing to Promote Financial Inclusion: Evidence from an Experiment in Nigeria	Nigeria	Lottery incentive scheme using mass media and social networks
Financial education and training		
Financial Education and Behavior Formation: Large-Scale Experimental Evidence from Brazil	Brazil	High school curriculum
The Impact of Financial Literacy Training for Migrants: Evidence from Australia and New Zealand	Australia & New Zealand	Group-based financial literacy seminar
Social Networks, Financial Literacy, and Index Insurance: Evidence from a Randomized Experiment in Kenya	Kenya	Comic books
The Impact of Financial Education on Financial Knowledge, Behavior, and Outcomes: Evidence from a Randomized Experiment in South Africa	South Africa	Group-based financial literacy seminar
The Impact of Financial Education and Learning-by-Doing on Household Investment Behavior: Evidence from Brazil	Brazil	Online stock market simulator
Financial Development and the Psychology of Savings Field Experiments in Rural Malawi	Malawi	Financial assistance and education
The Impact and Network Effects of Financial Management and Vocational Training in Informal Industrial Clusters: Evidence from Uganda	Uganda	Vocational training and financial capability training
Understanding Financial Capability through Financial Diaries and In-Depth Interviews in Uganda	Uganda	Financial capability training
Increasing the Impact of Conditional Cash Transfer Programs through Financial Literacy in the Dominican Republic	Dominican Republic	Financial capability training
Financial products and services		
The Role of Financial Access, Knowledge, and Service Delivery in Savings Behavior: Evidence from a Randomized Experiment in India	India	Opening of saving account; financial capability training
Does Financial Education Affect Savings Behavior? Evidence from a Randomized Experiment among Low-Income Clients of Branchless Banking in India	India	Bank cards; financial capability training
Evaluating the Effectiveness of Loan Disclosure Reforms on Consumer Understanding and Financial Decision Making: Evidence from Mexico	Mexico	Government-mandated product disclosure formats; phone and SMS financial capability counseling

illiterate; thus, such approaches may be particularly appropriate in lower-income countries.

RTF projects are also exploring innovations to more traditional forms of financial education. School children or new entrants to the labor force have long been an important group for financial education, in particular because financial capability may be inter-generationally transmitted and because children/youth may demonstrate the highest willingness and ability to learn. Furthermore, young adults are the most able to take advantage of the benefits of long-term planning. Interventions that aim to redress imbalances in the population are therefore best targeted at this group. School-based programs are useful models particularly in middle-income countries where school enrollment at post-primary levels is high. Two RTF projects in Brazil and Kenya explore new approaches to school-based financial education for high school students. In Brazil, the pedagogical innovation uses a purposively developed textbook that integrates financial capability into multiple aspects of the regular high school curriculum and also conducts parent workshops. In Kenya, another RTF project is delivering financial education in schools using comic books, in the context of a comic book that is already popular with the target population. A second delivery method is the more traditional classroom instruction at after school clubs. The study intends to also incorporate radio and social networking tools as a comparative element.

Other RTF projects are addressing the opportunities and needs that arise from recent trends in developing country financial markets. As suggested by the introduction to this chapter, a large amount of financial activity in the developing world is related to payments from governments to citizens and to remittances from migrant workers, and a growing share of it takes place through mobile banking and electronic payment cards. While the scope of government payment schemes is extremely wide, examples that naturally provide financial education opportunities include social payments, conditional cash transfers (CCTs), and pension distributions, all of which provide opportunities to bring individuals into the formal banking system and increase their understanding of financial issues. Government payment schemes also offer opportunities to combine education with incentives and benefits to encourage behavioral change. An RTF program is exploring this approach in the Dominican Republic by providing financial training alongside the Solidaridad CCT program. The evaluation will assess whether CCTs can be leveraged to include financial literacy education and access to financial services to effectively impact the productive assets, financial knowledge, behaviors, and outcomes of the poor. It also aims at evaluating whether increased literacy and access can themselves increase the impacts of the cash transfers and conditionality of the CCT program. In India, RTF-supported researchers are conducting an impact evaluation of a financial inclusion and literacy impact intervention delivered through “doorstep” banking in India. Another RTF project aims to estimate the causal impact of financial literacy training for migrant workers on their

remitting behavior, using a randomized trial in Australia and New Zealand, two countries with high shares of migrants from countries that depend heavily on remittances. Finally, one RTF program addresses the role of financial education in mitigating the lack of risk management noted among agricultural workers, evaluating the use of comic books to promote rainfall insurance among farmers in Kenya.

Taken together, the RTF programs also address a critical shortcoming in the literature—the lack of comparability across studies that examine only single interventions, making it difficult to fairly assess alternative interventions or programs both in terms of the size of effects and cost-effectiveness. Some programs are mindfully designed to leverage comparisons against other existing work: For instance, one study of financial literacy training among burial society groups and borrowing groups of the Women's Development Bank in South Africa is explicitly designed to draw meaningful cross-country comparisons with a similar study being conducted in India. Multiple RTF programs are conducting evaluations that do not simply test one financial capability intervention alone, but instead compare competing interventions in the same setting or the same intervention in different countries with comparable measures and protocols. For instance, in addition to the studies previously mentioned, in Mexico, RTF researchers will evaluate the benefits of improved disclosures and education about credit and savings products versus SMS messaging and incentives to contact a consumer hotline, analyzing a full spectrum of approaches to consumer protection. In the Dominican Republic, the RTF CCT evaluation includes a comparative examination of the impact of alternative combinations of interventions—transfers paid through a debit card, transfers paid with a bank account, basic training, in-depth financial literacy, opening of savings accounts, and access to a credit line.

The likely contributions of these evaluations go well beyond enriching our understanding of financial capability in the developing world, given the scarcity of current evaluation anywhere.

1.4 OVERVIEW OF THE TOOLKIT

The Toolkit is presented as a series of short chapters. Although it is important to be familiar with all parts of the monitoring and evaluation (M&E) process, it is not necessary to read the guide from beginning to end. Instead, each chapter is conceived as a standalone chapter that can be read independently of the others, according to each reader's needs. Each chapter includes a section on further reading for readers who would like to investigate any topic in more depth.

This Toolkit is intended to be a practical, hands-on guide to designing, conducting, and analyzing financial capability evaluations, with a focus on doing so in LMICs. The Toolkit covers a wide range of material on how to design, conduct and analyze evalu-

ations, material that is spread out over the 13 chapters that follow. The chapters are contained within four overarching parts: Setting the Stage for M&E: Understanding the M&E Process and Concepts (chapters 2–3); Conducting M&E for Financial Capability Programs (chapters 4–7); Collecting and Analyzing M&E Data for Financial Capability Programs (chapters 8–10); and Other Issues in Conducting M&E for Financial Capability Programs (chapters 11–14).

While all the chapters provide useful guidance, we recognize that some information is more useful for certain audiences than others. We have identified four key audiences for the Toolkit: policy makers, program staff, researchers, and evaluation experts. We anticipate that policy makers, who may be commissioning financial capability programs and/or evaluations, are most interested in a high-level overview of monitoring and evaluation. Program staff who are involved in the day-to-day operations of a financial capability program, will be interested in details of conducting monitoring and evaluation, as well as practical guidance. Researchers and evaluation experts who will be designing and conducting the evaluation and analyzing data, will be most interested in details of conducting monitoring and evaluation and data analysis. Table 1.2 provides a breakdown, by part and chapter, showing which chapters are most relevant for which audiences.

The Toolkit also includes some appendixes that provide additional resources for readers who desire more detail. Appendix A describes the RTF Financial Capability Evaluation projects. Appendix B provides technical details on the methodology described in chapter 6. Appendix C (available in the online version of this report and accessible at www.finlitedu.org) reproduces a report from the World Bank RTF Financial Capability Measurement Program team describing the development of their Financial Capability Survey. Appendix D (available in the online version of this report and accessible at www.finlitedu.org) reproduces a report on the Global Financial Inclusion database. Appendix E (available in the online version of this report and accessible at www.finlitedu.org) is a collection of data collection instruments used for collecting data on financial capability.

TABLE 1.2 A READER'S GUIDE TO THE TOOLKIT

CHAPTER NO./DESCRIPTION	POLICY MAKERS	PROGRAM STAFF	RESEARCHERS & EVALUATION EXPERTS
Part I: Setting the Stage for M&E: Understanding the M&E Process and Concepts			
2. MONITORING AND EVALUATION: lays out M&E concepts and the importance of M&E to financial capability programs	X	X	
3. SETTING THE STAGE: lays out and illustrates the fundamental process of M&E efforts in financial capability programs	X	X	
Part II: Conducting M&E for Financial Capability Programs			
4. MONITORING: lays out why to monitoring and the key elements of a monitoring system for financial capability programs		X	X
5. PROCESS EVALUATION: lays out why to do process evaluation and the logical steps of a process evaluation for financial capability programs		X	X
6. IMPACT EVALUATION: lays out key terms and concepts related to impact evaluations and introduces quantitative methods available for conducting impact evaluations for financial capability programs		X	X
7. PUTTING IT ALL TOGETHER: lays out a comprehensive or "mixed methods" approach that incorporates monitoring and process and impact evaluations for financial capability programs		X	X
Part III: Collecting and Analyzing M&E Data for Financial Capability Programs			
8. DATA COLLECTION METHODS: lays out types of quantitative and qualitative data collection methods and their uses in financial capability program evaluation			X
9. THE PROCESS OF COLLECTING DATA: PRACTICAL GUIDANCE: lays out some practical techniques in using qualitative and quantitative data collection methods for financial capability programs		X	X
10. ANALYZING QUANTITATIVE AND QUALITATIVE DATA: lays out basic concepts in analyzing quantitative and qualitative data for financial capability program evaluations			X
Part IV: Other Issues in Conducting M&E for Financial Capability Programs			
11. COST ANALYSIS: WEIGHING PROGRAM COSTS AND BENEFITS: lays out the steps in cost analysis and provides simple illustrations from financial capability program evaluations			X
12. IMPLEMENTING THE EVALUATION: lays out the practical aspects of building an actionable plan for conducting financial capability programs		X	X
13. ETHICAL CONSIDERATIONS: lays out the ethical issues that may arise in designing a financial capability program evaluation and in collecting and analyzing data from it	X	X	X
14. DOCUMENTING AND COMMUNICATING RESULTS: lays out the importance of documenting and communicating financial capability program evaluation results and how to do so effectively		X	X

FURTHER READING

Bernheim, B. D., D. M. Garrett, and D. M. Maki. 2001. "Education and Saving: The Long-term Effects of High School Financial Curriculum Mandates." *Journal of Public Economics* 80 (3): 435–65.

Bruhn, M., and B. Zia. 2011. "Stimulating Managerial Capital in Emerging Markets: The Impact of Business and Financial Literacy for Young Entrepreneurs." Policy Research Working Paper Series WPS5642, World Bank, Washington, DC. As of March 20, 2012: http://www-wds.worldbank.org/external/default/WDSContentServer/IW3P/IB/2011/04/27/000158349_20110427082512/Rendered/PDF/WPS5642.pdf.

Carpena, F., S. Cole, J. Shapiro, and B. Zia. 2011. "Unpacking the Causal Chain of Financial Literacy." Policy Research Working Paper WPS5798, World Bank, Washington, DC. As of April 1, 2012: http://www-wds.worldbank.org/servlet/WDSContentServer/WDSP/IB/2011/09/19/000158349_20110919154530/Rendered/PDF/WPS5798.pdf.

Cole, S., T. Sampson, and B. Zia. 2011. "Prices or Knowledge? What Drives Demand for Financial Services in Emerging Markets?" *Journal of Finance* 66 (6): 1933–67.

Cole, S., and G. K. Shastry. 2009. "Smart Money: The Effect of Education, Cognitive Ability, and Financial Literacy on Financial Market Participation." Harvard Business School, Working Paper.

Collins, J. M., and C. O'Rourke. 2009. "Evaluating Financial Education and Counseling: A Review of the Literature." As of March 2, 2010: http://ssrn.com/abstract=1529422.

Collins, J. M., and C. O'Rourke. 2011. "Homeownership Education and Counseling: Do We Know What Works?" Research Institute for Housing America Special Report, Mortgage Bankers Association. As of April 1, 2012: http://www.housingamerica.org/RIHA/RIHA/Publications/76378_10554_Research_RIHA_Collins_Report.pdf.

Drexler, A., G. Fischer, and A. Schoar. 2011. *Keeping it Simple: Financial Literacy and Rules of Thumb*, MIT Working Paper.

Duflo, E., and E. Saez. 2003. "The Role of Information and Social Interactions in Retirement Plan Decisions: Evidence From a Randomized Experiment." *Quarterly Journal of Economics* 118 (3): 815–42.

Fox, J. J., and S. Bartholomae. 2008. "Financial Education and Program Evaluation." In J. J. Xiao, ed., *Handbook of Consumer Finance Research*, New York: Springer, pp. 47–68.

Giné, X., and G. Mansuri. 2011. *Together We Will: Experimental Evidence on Female Voting Behavior in Pakistan*, Policy Research Working Paper Series WPS5692, World Bank, Washington, DC.

Hogarth, J. M. 2006. *Financial Education and Economic Development*. Presented at the G8 International Conference on Improving Financial Literacy, Moscow, Russian Federation, 29 November 2006. As of April 1, 2012: http://www.oecd.org/dataoecd/20/50/37742200.pdf.

Holzmann, R., F. Mulaj, and V. Perotti. 2013. *Financial Literacy and Education in Low- and Middle-Income Countries: Measurement and Effectiveness. A Report on the Research Program and Knowledge Derived from the Russia Financial Literacy and Education Trust Fund*. Washington, DC: World Bank.

Karlan, D., and M. Valdivia. Forthcoming. "Teaching Entrepreneurship: Impact of Business Training on Microfinance Clients and Institutions." *Review of Economics and Statistics*.

Kempson, E., and A. Atkinson. 2009. *Measuring Levels of Financial Literacy at an International Level.* Paris: OECD.

Kempson, E, S. Collard, and N. Moore. (2005), *Measuring Financial Capability: An Exploratory Study*. London, Financial Services Authority.

Lusardi, A. 2004. "Savings and the Effectiveness of Financial Education." In Olivia S. Mitchell and Stephen Utkus, eds., *Pension Design and Structure: New Lessons from Behavioral Finance*, Oxford: Oxford University Press, pp. 157–84.

Lusardi, A., and O. S. Mitchell. 2007. "Financial Literacy and Retirement Preparedness: Evidence and Implications for Financial Education." *Business Economics* b42 (1): 35–44.

Lusardi, A., and O. S. Mitchell. 2008. "Planning and Financial Literacy: How Do Women Fare?" *American Economic Review: Papers and Proceedings* 98 (2): 413–17.

Lusardi, A., Robert L. Clark, Jonathan Fox, John Grable, and Ed Taylor. 2010. "Promising Learning Strategies, Interventions, and Delivery Methods." In *Financial Literacy Education: What techniques, venues, tactics, mechanisms, etc., Show the Most Promise to Promote and Achieve Financial Well-being?* As of April 1, 2012: http://www.nefe.org/LinkClick.aspx?fileticket=4x7s1YB3-V4%3d&tabid=934.

Lyons, A. C. 2005. "Financial Education and Program Evaluation: Challenges and Potentials for Financial Professionals." *Journal of Personal Finance* 4 (4): 56–68.

Lyons, A. C., L. Palmer, K. S. U. Jayaratne, and E. Scherpf. 2006. "Are We Making the Grade? A National Overview of Financial Education and Program Evaluation." *Journal of Consumer Affairs* 40 (2): 208–35.

Lyons, A. C., M. Rachlis, M. Staten, and J. J. Xiao. 2006. "Translating Financial Education into Knowledge and Behavior Change." *Consumer Interests Annual* 52. As of September 12, 2007: http://www.consumerinterests.org/i4a/pages/Index.cfm?pageid=4145.

Mandell, L. 2009. "Financial Education in High School." In Annamaria Lusardi, ed., *Overcoming the Saving Slump: How to Increase the Effectiveness of Financial Education and Saving Programs*, 257–79. Chicago: University of Chicago Press.

Martin, M. 2007. "A Literature Review on the Effectiveness of Financial Education." Working Paper No. 07-03, Federal Reserve Bank of Richmond, Richmond, VA. As of March 2, 2010: http://www.richmondfed.org/publications/research/working_papers/2007/wp_07-3.cfm.

Noctor, M., S. Stoney, and R. Stradling. 1992. *Financial Literacy*. Report prepared for the National Westminster Bank, London.

Schuchardt, J., S. D. Hanna, T. K. Hira, A. C. Lyons, L. Palmer, and J. J. Xiao. 2009. "Financial Literacy and Education Research Priorities." *Journal of Financial Counseling and Planning* 20 (1): 84–95.

Zia, B. 2011. "The Impact of Financial Literacy Around the World: Motivation, Existing Evidence, and Upcoming Evaluations." DIME Conference Presentation, June 2011. As of April 1, 2012: http://siteresources.worldbank.org/INTDEVIMPEVAINI/Resources/Zia-ImpactofFinancialLiteracyAroundtheWorld.pdf.

PART I

Setting the stage for M&E: understanding the M&E process and concepts

CHAPTER 2

Monitoring and evaluation

During or after a financial capability program, those who manage it—along with policy makers and other stakeholders—will want to know whether the program is delivering what was hoped for, what is working well and what is working less well, and what can be done to maximize success. In fact, having some mechanism in place from the outset to do this may be a condition of program funding.

These needs are typically met through **monitoring and evaluation**, often called M&E. M&E is not a single process; rather, it is two distinctly different processes, although they may be linked to one another. **Monitoring** is the regular collection and analysis of information on a program or service, with the ultimate goal of assessing the program's progress against its plans and intended milestones. **Evaluation**, in contrast, is a periodic or one-time detailed analysis of program performance against predetermined aims and objectives. As noted, the two are separate, but they may be linked because monitoring data may be used as an input in the evaluation, alongside other information.

Monitoring and evaluation are two distinctly different processes, although they may be linked to one another.

This chapter introduces M&E concepts to help facilitate decisions on the best approach for evaluators to take in assessing a program—in particular, the kinds of evaluations that can be done. The chapters that follow will go into much greater detail about the concepts briefly introduced here.

2.1 WHAT IS MONITORING?

Monitoring involves regularly collecting, analyzing, and reviewing information about how a financial capability program is operating. Monitoring is primarily descriptive and can be used to gain rapid, real-time insights into the program as it develops and on the progress it is making toward meeting work plans and milestones. It can also be used to track the use of the program's budget and to identify any emerging threats to the program's success. A typical checklist for a monitoring effort might include the kinds of questions shown in the checklist in table 2.1.

Where does the information to answer these kinds of questions come from? Monitoring information may be collected and brought together from a range of sources,

TABLE 2.1 SAMPLE MONITORING CHECKLIST

TYPICAL QUESTIONS	EXAMPLES
Have the resources been used as intended? Are the quantity and quality of the program inputs as you had planned?	▪ Were materials for workshops sufficient for the number of participants? ▪ Were they clear and accessible to the target audience?
Do the actual activities undertaken correspond to those that were planned?	▪ If six workshops at each of 10 locations were planned, did all these actually take place as planned?
Have the expected products and services been delivered to the planned standard?	▪ Did the workshops cover the intended subjects and were the planned supporting materials for participants produced and distributed? ▪ How many people participated in the workshops?
Has the target audience been reached?	▪ If the program delivery was spread across more than one day, how many people participated in all of the sessions? ▪ Were they the intended beneficiaries?
Have the intended outcomes been achieved?	▪ If the aim of the program was to encourage regular saving into a particular type of account, how many participants have begun to do so?

including both the organization planning the program and those charged with delivering it. And it may also involve third parties, particularly if you want to monitor changes in participants' behavior.

For more information on how to design and set up a monitoring program see chapter 4.

2.2 WHAT IS EVALUATION?

Formative evaluation provides information on program delivery.

Summative evaluation focuses on the outcomes and impact of delivered program services.

Monitoring, if done well, provides good information on how the program is operating and being delivered to its intended beneficiaries, but the reporting and analysis is primarily descriptive in nature. Evaluation, in contrast, involves judgment. Evaluation assesses how well the program is performing and how much success it is having in achieving its intended outcomes,

Evaluations are conducted at strategic moments in the program cycle, such as a short period after its inception, or when it has been operating for some time and achieved maturity. If an evaluation is conducted during the start or in the midst of the program, it is typically referred to as a **formative** evaluation; its primary goal is to provide feedback on the program's delivery to inform the evolution of the program itself. An evaluation that is conducted after the program has been established or even concluded is called a **summative** evaluation. It may focus on how well the program met its objectives with respect to deliverables, but also on the outcomes

and impact of the services delivered. Data collected at inception can provide baseline information for both formative and summative evaluations.

There are two main types of evaluation: **process evaluations** and **impact evaluations**. Although the same information sources can be designed to provide inputs for both process and impact evaluations, each is designed to answer different types of questions. A robust overall program evaluation would include both.

There are two main types of evaluations: process and impact evaluations. Each is designed to answer different types of questions. Process evaluations address implementation, while impact evaluations address the program's effects.

2.2.1 Process evaluation

Process evaluations address a broad range of questions about the nature of a financial capability program's development and implementation, including its relevance, efficiency, and effectiveness:

- Were the program's goals and objectives appropriate for the specific context?
- Was the program implemented as intended? What lessons can be learned about optimizing implementation?
- Did the inputs and activities achieve the project goals?
- Did the program reach the intended beneficiaries? If it did not, then why not? And what lessons can be learned?
- Was the quality of the project implementation adequate? What lessons can be learned about improving quality?
- Were the intended outputs delivered and objectives achieved? If they were not, then why not? And what lessons can be learned?
- Did the program have any unplanned or unintended effects?

As these questions indicate, there is an emphasis throughout a process analysis on lessons learned; the intent is to understand what worked and did not work as the program and its services were delivered. As a result, it is necessary for evaluators conducting a process evaluation to typically collect information from a range of stakeholders, including the people responsible for planning and delivering the program, the beneficiaries (and often intended beneficiaries who do not use it), and others. In doing so, evaluators use a combination of data collection methods, both quantitative (such as surveys) and qualitative (such as in-depth interviews, focus groups, mystery shopping, observation, and site visits). The process evaluation may also use monitoring information as an input. These methods of data collection are described in more detail in chapter 8.

Process evaluations can be either formative or summative—or they can cover both. In other words, they can contribute to the design and delivery of the program at key milestone points as it is being developed (formative) and assess it once it has been

established (summative). This last point is an important one. Everyone wants results as quickly as possible, but a summative process evaluation conducted too early in a program will rarely be money well spent. It is important to wait until the initial problems that are encountered in any new program have been sorted out and it is running smoothly.

2.2.2 Impact evaluation

Unlike a process evaluation, which is focused on input, activities, and outputs relative to goals and objectives, an impact evaluation is focused on outcomes. In other words, it measures the causal effect of the program and its services on the observed outcomes—the impact. In the case of a financial capability program, impact will refer particularly to changes in behavior among the beneficiaries, but also to changes in knowledge, skills, and attitudes related to financial decision making that underlie behavioral changes. So, for example, the impact of a financial capability program could be measured as the change in the proportion of people saving regularly or who plan how they will spend their money.

Evaluations that only compare participants before and after program participation or that only compare participants to nonparticipants after the program generally cannot accurately assess program impacts.

When we talk about the "causal effects" of a program, we are interested in the outcomes that can be **directly attributed** to the financial capability program, and that could not have been caused by anything else.

Not all approaches that evaluators commonly use enable one to accurately measure the outcomes directly attributable to a program. Put another way, the impact of a program is the difference between the observed outcomes and what the outcomes would have been if the program had not taken place. In technical terms, this is known as "the **counterfactual**." A common approach taken by many evaluations is to compare measures of participants' behaviors, knowledge, skills, or attitudes before and after they take part in a program. These **prepost** evaluations may even report participants' own perceptions of how they have changed in these respects. Unfortunately, evaluators cannot credibly say whether any of the observed outcomes would or would not have happened in the absence of the program, especially if other programs or campaigns that could affect the target outcome happen at the same time.

Another approach taken by some evaluators is to compare people who have participated in a program with those who did not. However, in this case we still cannot be sure that differences in the outcomes of interest between these two groups are products of the program. The problem here is that there may be important differences between these two groups that will affect the outcomes observed. It is important to note that this is true even if the program itself does not impose restrictions on participation: individuals themselves choose to join or not based on differences in their own characteristics, leading to **self-selection.**

For example, suppose that an NGO decides to hold a series of financial-planning workshops in a large city to encourage savings, and tracks the savings of participants before enrollment and a year later. The results show that participant savings rates have increased over time! However, at the same time, during the year, more bank branches have opened in the city, making it easier to open savings accounts. Regional banks have also run widely noticed advertisements encouraging savings that may also play a significant role in influencing savings rates in the whole area. As a result, the observed changes in the participant group may not be due to the NGO's workshop, as it is difficult, if not impossible, to rule out a number of alternative and plausible explanations. Comparing participants to nonparticipants after the program is also problematic. The people who sign up for a financial planning workshop on saving could be those most motivated to save, while those who do not sign up have lower levels of motivation. Participants in the workshop might then have higher levels of savings relative to nonparticipants, with or without the workshop. This means that any differences observed between the two groups may be the result of preexisting motivations and, thus, cannot be directly attributed to the workshop.

A good impact evaluation, in contrast to these two approaches, needs to be able to "tease out" causal impact that is directly attributable to the program in question. Doing so requires comparing program participants with a group of people who are identical in all respects except that they did not participate in the program. There are various ways of selecting this group of people, who are known as the **control** or **comparison** group depending on how they have been selected. This is described in more detail in chapter 6.

An impact evaluation normally forms part of a summative evaluation and, like the process evaluation described above, should not be carried out too soon in the life of a new program. It is important that any initial problems are ironed out because these will almost certainly have affected outcomes. It will generally use an **experimental** or **quasi-experimental** approach (described in chapter 6) and be based on quantitative data collection, such as a survey of the participants and the control or comparison group. The same survey can also be used to collect information for a process evaluation. There may also be occasions when monitoring information is used to measure outcomes, especially those relating to changes in behavior. Examples of such information for financial capability programs include the opening of bank accounts, changes in savings patterns, and changes in patterns of repayment of credit commitments.

However, measuring the magnitude of impact is only one part of an impact evaluation. It is just as important to understand how the program achieved this impact, and what explains variations in the observed outcomes. For example, a well-designed financial capability program to encourage prompt repayment of credit commitments

may have been undermined by a drought or some other factor that has affected people's incomes. Alternatively, changes in personal circumstances for certain groups (for example, new parents or those experiencing separation from a partner) may have affected their ability to repay. How the program was designed or the services delivered may also affect the impact. Exploring links to process analysis and qualitative research on outcomes can provide the necessary context. For instance, an understanding of how influences such as these have affected outcomes can be provided by a small number of in-depth interviews—which, in turn, might suggest fruitful lines of analysis of the survey data, as well as linkages to the process analysis.

For more information on how to design and carry out an impact evaluation, see chapter 6.

2.2.3 Cost analyses

Process and impact evaluations are important to help decision makers understand whether a program delivered what was promised and how well it performed relative to other programs in terms of benefits to participants. However, this is not all: funders often face many competing needs for resources. Those who are supporting a program financially need to understand whether the program's delivery and performance provide more value for their money relative to other alternative programs or whether the resources spent would be put to better use elsewhere. Evaluators and other stakeholders (such as funders) and policy makers therefore often also need to assess the delivery and achievements of a program against the costs incurred in providing those, relative to other alternatives.

Cost-benefit: How do total costs of implementing the program stack up against the total expected benefits?

Cost-effectiveness: How does the cost per output (outcome) compare with that of similar programs in achieving that output (outcome)?

Cost-consequences: What are all the relevant costs and benefits to a program?

There are three tools for doing this, all of which draw inputs from both process and impact evaluations.

A **cost-benefit** analysis weighs the total costs incurred by a financial capability program against the total benefits (outcomes), measured in monetary terms, to come up with the net benefit of undertaking the program. Assigning a monetary value to a benefit or outcome is often quite difficult to do, but one example might be a program designed to encourage people to save up rather than use credit to buy goods or services, where the net financial savings to the individual can be calculated. A cost-benefit analysis is often used in combination with an impact evaluation and together they can provide very powerful evidence for investing in a particular program—or not doing so!

A **cost-effectiveness** analysis measures the cost per output or outcome (e.g., $300 per savings account opened) in a financial capability program and usually compares this cost with other similar programs or other ways of achieving the same outcome. As such, a cost-effectiveness analysis tells evaluators what the program got for each

dollar spent and whether this is commensurate with expectations. Such an analysis will also enable stakeholders and policy makers to decide whether the program, as currently designed and delivered, represents value for money. And, if carefully designed, it can determine whether, say, only a slightly lower level of effectiveness could be achieved for significantly less money. It is particularly valuable where a program is delivered in a number of different ways (e.g., debt repayment information delivered through a public service announcement compared to the same information delivered through a comic book), or for benchmarking a program against others because the cost-effectiveness of the different models can be compared.

A **cost-consequences** analysis enumerates and characterizes all the relevant costs and benefits for a program. A cost-consequences analysis uses quantitative (monetary) costs and benefits where possible, and qualitative descriptions otherwise. Tabulating costs and benefits in a cost-consequences analysis is an easy way to help decision makers understand the bigger picture of costs and benefits.

For more information on cost-benefit, cost-effectiveness, and cost-consequences analyses, see chapter 11.

2.3 DECIDING WHEN AND HOW TO UNDERTAKE M&E

So far, we have talked about what M&E activities are and what types of evaluations there are. Ideally, one would decide at the outset of planning a program that an evaluation will be carried out—this means the evaluation would be "prospective." If evaluators are involved at the very beginning of a program or as part of a pilot test, developing the M&E framework and the program planning can go hand-in-hand. Not surprisingly, incorporating M&E **prospectively** greatly enhances the quality of evaluation for several reasons. Beyond the value of providing immediate feedback on the program itself, planning from the start will ensure that: the goals set for the program are appropriate and measurable; that the data needed for the evaluation are collected from the outset, which is essential if baseline data are required for comparison purposes; and that the evaluation is viewed as ongoing and constructive, which should increase the likelihood that the results are credible. It may also result in cost efficiencies, because M&E activities can then be structured to take advantage of program infrastructure in a manner appropriate to its scale and resources. Finally, it will, of course, be very difficult to assess impact if the evaluation is not planned from the outset.

Incorporating process and evaluation planning prospectively greatly enhances the quality of the evaluation.

Of course, not everything follows the ideal. Often, planning turns out to be **retrospective**—this means that program officials or external stakeholders decide to evaluate an ongoing or even a completed program after the fact. Not having even a formal monitoring system from its inception will complicate things further. While

retrospective evaluation is not the ideal, it does not mean it is not doable. What it does mean is that retrospective evaluations must be carefully deliberated and managed to ensure both that it is feasible and that it can provide answers to the key questions that need answering. In particular, the ability to attribute effects to a financial capability program—or conduct impact evaluations—may be limited.

While we often think of research as being conducted by trained evaluators independent of those being evaluated, that is not always the case. Many research activities are "participatory" in nature, where evaluators work directly with stakeholders to seek change and, thus, involve them evaluation processes or inquiry. This is obviously the case when evaluation is prospective, but is also true for retrospective evaluation.

For evaluations to be successful, enough resources and training must be provided to develop staff evaluation skills and capabilities, and the program must be properly grounded so that the evaluation can be done credibly.

Research can be participatory in many different ways—structured, for example, around group activities or stakeholder interviews that generate qualitative feedback into both qualitative and quantitative aspects of evaluation. In program evaluations, participatory research allows not just the evaluators but also program staff, beneficiaries, and other stakeholders to provide input, for instance, into research questions, the types of outcomes to examine, appropriate ways to conduct qualitative research, the meaning of certain findings, and the dissemination of results to the rest of the community and other interested parties.

Taking a participatory approach can help to make an evaluation design more robust, and its findings more effective. It is important however to understand that the use of participatory research methods should be balanced against a commitment to objectivity in carrying out the evaluation as a whole. The evaluator's responsibility is to ensure that stakeholder input is balanced and informs but does not dictate the evaluation, using his or her judgment appropriately and sensitively.

For example, if a rural bank aims to evaluate a program, early evaluation activities can include village meetings to identify key problems. The community can be involved in the collective creation of maps of local areas indicating where people's fields, houses with and without electricity, and any number of other features of interest are located. Evaluation findings may be discussed in the community, and disseminated at both the grassroots and policy level.

2.4 MAKING THE DECISION TO EVALUATE

Evaluation activities can be costly, both in terms of time and resources. To that end, decision makers may often wonder how much effort to devote to evaluation in addition to regular program monitoring, and whether or not to evaluate at all.

2.4.1 What type of evaluation is desirable?

At the most fundamental level, evaluation is necessary to understand whether and how a program is working, in addition to keeping track of its activities through monitoring. The key question is not **whether** a particular program should be evaluated, but **how**. The scope and scale of evaluation should be commensurate with the scale of the program and/or its larger strategic objectives.

A strong strategic case for evaluation can be made if any of the following are true of the program:

- Is the program in question innovative? Has the impact of similar programs never been tested?
- Is it a pilot program that will later be under consideration for expansion?
- Will the program results be useful in helping design and deliver the program in the future? Is it replicable and will others want to learn from it?
- Does the program involve significant resource allocation?
- Can strong and sustained stakeholder commitment be expected if the results are positive?

If the answers are "no," then careful thought should be given to the scale and scope of evaluation; it does not mean that an evaluation is not needed, but rather that a more modest approach may be called for. For example, process evaluations are generally within the scope of most program budgets, but impact evaluations and cost-analyses analyses often need to be used more selectively, largely because of the resource and information requirements involved.

A second and slightly different consideration is that doing evaluations—and doing them right—is very challenging. This means a number of additional questions should be asked about the program capacity:

- Are there enough internal and external resources to collect, manage, and analyze the data needed for the evaluation?
- Does the staff that will carry out the evaluation have the skill set and capability to do so?

If the answers to the questions above are "no," that does not mean that an evaluation cannot be done. But it does mean that effort must be taken to the turn the "no's" into "yes's." In some cases, the need is for finding resources (both money and staff); in other cases, the need is for training and staff development to get the needed skills and capabilities to conduct evaluations.

2.4.2 What type of evaluation is possible?

A second consideration is to understand what type of evaluation design is conceptually feasible, in addition to being desirable. This can mean different things depending on the type of evaluation planned. As previously noted, for a summative evaluation, this means determining whether the program is sufficiently developed and stable enough for such an evaluation. Alternatively, one should consider that while process evaluation may be implementable under many circumstances, it may not always be possible to conduct a rigorous impact evaluation. Programs may have concluded or have been implemented in such a way as to make it impossible to establish a valid control or comparison group. For example, if evaluators go ahead with a before-and-after survey with participants or an impact evaluation with participants and nonparticipants who have not been rigorously selected, such an evaluation will lead to results that can be dangerously misleading, especially when its limitations are not well understood. Depending on the objectives of the evaluation, a better use of resources may be to focus on the process evaluation, combined with in-depth interviews to understand the role that the program played in bringing about the outcomes of interest and how it related to other influences on those outcomes.

2.4.3 An evaluability assessment

It is often useful to think about the desirability and feasibility of evaluation in a systematic way, by conducting a preliminary assessment. A reasonable **minimum** strategy for an evaluability assessment of what type of evaluation to conduct should include:

- A detailed **desk review** of relevant documents to study the history, design and current characteristics of the program.
- **Interviews** with key stakeholders including program staff to understand the above, but also their needs and commitment toward evaluation, feedback on the program itself and its operation, as well as their potential to contribute to the evaluation (e.g., funding, staff time, access to information).
- **Site visits** and **observation** of activities to verify existence of capacity for evaluation and to watch the program in action.

The final judgment of evaluability should be led by the likely lead evaluator but involve input from key stakeholders including program planners and implementation staff. The results of such assessment should be reasonable consensus on a shortlist of potential evaluation questions and a preliminary understanding of the possible scope, methods and design of the evaluation to be deployed.

KEY POINTS

Conducting M&E activities is a key part of implementing a financial capability program for any number of reasons. It can establish that the financial capability program is on the right track, that it is being implemented as intended, and that it is having the expected impact. And conducting M&E activities may be a condition of getting funding for the financial capability program in the first place.

Table 2.2 lays out the goals, example questions, and example activities, as well as the timing and frequency of monitoring and the types of evaluations. Remember, these are examples of questions and activities, not a comprehensive list.

TABLE 2.2 MONITORING, PROCESS EVALUATION, AND IMPACT EVALUATION

GOAL	EXAMPLE QUESTIONS	EXAMPLE ACTIVITIES	TIMING/ FREQUENCY
Monitoring			
Track and manage program resources	▪ Have the resources been used as planned? ▪ Were the activities undertaken as planned? ▪ Have the activities been delivered to the standard envisaged? ▪ Has the intended target audience been reached? ▪ Have the intended outcomes been achieved?	▪ Count inputs and outputs ▪ Keep cost accounts	Throughout the program
Process evaluation			
Document and assess implementation, operation, and outcomes against the goals and objectives of the program	▪ Were the goals and objectives appropriate? ▪ Was the program implemented as intended? ▪ Were the goals and objectives met? ▪ Was the quality of the implementation adequate? ▪ Did the program reach the intended beneficiaries? ▪ Were the intended outcomes achieved? If not, why not? And what lessons can be learned? ▪ Did the program have unplanned or unintended effects?	▪ Document administrative process ▪ Review all monitoring data for trends and patterns ▪ Interview people involved in the planning and delivery of the program ▪ Interview beneficiaries in a survey or in-depth interviews ▪ Directly observe the program	Ad hoc, at inception at an interim point, and when service has been established
Impact evaluation			
Measure the causal effect of program	▪ What impact has the program had? ▪ Does the impact justify the cost?	▪ Conduct baseline and follow-up survey on treatment and control participants	Ad hoc, when service has been established

It is important not to carry out a summative evaluation (whether process or impact) too early.

As such, it is, as we discussed above, worth investing time and effort to decide what type of evaluation is most appropriate and whether it is feasible and likely to provide credible results. All told, an evaluability assessment should develop a preliminary but clear grasp of the program at the outset, including its goals and plans, and how much it is being carried out according to the original goals and plans. It should also provide an initial assessment of the justification for evaluation, its logistical feasibility, the potential usefulness and credibility of the results, and the strength of political and financial support. Finally, the assessment should attempt to characterize factors such as the quality of any monitoring data that has been collected, the type of program, and the nature of the intended impacts.

A systematic process of assessment can help to formalize understanding between an evaluation team and those involved in planning and delivering a program, to set priorities, to anticipate problems, and to set mutual expectations. It is useful to start this process by clarifying the program logic or theory of change, which is the subject of the next chapter.

FURTHER READING

Gertler, P. J., Martinez, S., Premand, P., Rawlings, L. B., and Vermeersch, C. M. 2011. "Impact Evaluation in Practice," World Bank Publications.

Kusek, J. Z., and Rist, R. C. 2004. "Ten Steps to a Results-Based Monitoring and Evaluation System: A Handbook for Development Practitioners," vol. 289, World Bank Publications.

Levinson, F. J., Rogers, B. L., Hicks, K. M., Schaetzel, T., Troy, L., and Young, C. 1999. "Monitoring and Evaluation of Nutrition Programs in Developing Countries," *Nutrition Reviews* 57 (5): 157–64.

UNPF (United Nations Population Fund). 2002. "Monitoring and Evaluation Toolkit for Program Managers," Office of Oversight and Evaluation. Available at http://www.unfpa.org/monitoring/toolkit.htm.

Gosling, L., Edwards, M. (1995), "Toolkits—A Practical Guide to Assessment, Monitoring, Review and Evaluation," Development Manual 5, Save the Children.

Bamberger, M., K. Mackay, and E. Ooi (2005), "Influential Evaluations: Detailed Case Studies," Operations Evaluation Department. Washington, DC: World Bank.

CHAPTER 3

Setting the stage

In the last chapter, we talked about the concepts and approach to conducting monitoring and evaluation (M&E) for a financial capability program. But regardless of whether evaluators will carry out monitoring, a process evaluation or an impact evaluation, or all of the above—all M&E efforts must start by formally laying out and providing context for the financial capability program goals and objectives. If we don't know what the program is trying to achieve (its goals) or how we plan to concretely achieve those goals (its objectives), then we can't expect to be able to determine if it's being carried out effectively (through a process analysis) or if it's having the desired impact (through an impact analysis).

This chapter lays out and illustrates this fundamental process of "setting the stage" for the M&E efforts that follow.

3.1 HOW SHOULD EVALUATORS GO ABOUT FORMALIZING PROGRAM GOALS AND OBJECTIVES?

Programs don't exist in a vacuum. A critical first step to formalizing the goals and objectives of any planned financial capability program is to step back and get a better understanding of the problem the program is trying to address and the potential ways to solve it. This "stepping back" process is often referred to as a **resources and needs assessment** or a general **situation analysis.**

Conducting situation analyses narrows down the specific problem to address in the proposed program, the target populations to focus on, and what makes sense in terms of desired outcomes.

Such analyses can include a literature review of existing research related to the substantive area of the planned program and/or its target population, a desk review of program documents for alternative programs addressing similar issues, surveys and/or key informant interviews with potential program beneficiaries and other stakeholders, and other types of relevant qualitative research. Situation analyses like these perform the valuable function of narrowing down the specific problem to be addressed, the target populations to focus on, and what makes the most sense in terms of desired outcomes to be measured.

In the previous chapter, we talked about prospective and retrospective studies. If the proposed financial capability program is being simultaneously developed with an evaluation from the get-go—a prospective study—then a situation analysis would be the first part of the planning stage and would serve as the basis for the planner to develop an overall program strategy. Such a strategy may entail expanding an existing program or best practice that came up in the situation analysis and adapting it to the context of the target population the intended program will serve. For retrospective studies, this situation analysis would build on the initial evaluability assessment to understand the nature and magnitude of the proposed program's constraints and to identify potential issues and subgroups of interest and programs to use as potential comparisons.

For example, a Russia Financial Literacy and Education Trust Fund (RTF) –supported pilot program in Nigeria carried out a situation analysis for the evaluation of a short film on financial literacy covering the following topics: (1) personal and household day-to-day financial management, (2) short- and long-term financial planning, (3) planning for unexpected needs, and (4) choosing different financial products. Evaluators were considering whether or not to expand the scope of evaluation, specifically whether to undertake a larger-scale effort to compare the impact of all the different platforms used for distribution. In their situation analysis, the evaluation team conducted key-informant interviews with the nongovernmental organization (NGO) that had developed the film and was coordinating its screenings. The aim of these interviews was to understand the NGO's rationale for developing the film and rolling out the intervention in the form of screenings, rather than DVD or CD distribution. According to the NGO, participants would be less likely to consider the content of the movie as important if they received a DVD for free than if they attended a screening at a public venue. These interviews indicated to the evaluation team that the evaluation should focus on the impact of the movie itself, rather than a comparative evaluation.

With a situation analysis completed, the next step is to clearly define the program goals and objectives. While they may sound like the same thing—and indeed are aligned and mutually reinforce each other—they are actually very different.

Goals are broad program aims. Objectives are best defined as operationalized goals that must be measurable, well defined, and realistic.

- **Goals** are **broad program aims** that are not necessarily expressed in concrete, quantifiable terms. Goals summarize the program's theory of change—what the program planners hope to attain—but do not provide guidance for meaningfully measuring the degree of a program's achievement.
- **Objectives** are best defined as **operationalized goals**; in other words, they are what you need to do in practice to actually achieve the broader program goals. Such objectives allow program evaluators to compare program accomplishments to initial expectations. What defines objectives? First, they must be **measurable and well defined** in terms of specific indicators, something

we discuss in more detail below. Second, they must be **realistic** given the resources and time frame at hand. For example, if an objective of a one-year financial capability program is to measure behavioral changes that will not show up until years after the program has ended, then the objective is not realistic in the time frame of the program. All this means that objectives must set out expected or desired magnitudes of change in specifically defined indicators over a definite period. All this sounds very quantitative, but it is important to understand that even objectives built around qualitative measures can be defined in measurable ways (as the examples to be discussed shortly will show).

What might some actual goals and objectives look like in a financial capability program? An example of clear program goals and objectives is provided by one of the RTF's pilot programs based in South Africa. This pilot program is embedding financial capability storylines in a soap opera called *Scandal!*. Financial capability storylines were developed by the show's production company and Ochre Media, along with the South African National Debt Mediation Association (NDMA) and a team of financial capability and social marketing experts.

The overall goal of the program was to increase financial capability among the audience. The main objectives of the program were to improve the knowledge, attitudes and behavior of low-income South Africans regarding financial decision making with a focus on debt management. Changes in knowledge, attitudes, and behaviors among the target population are being measured in an impact evaluation using various indicators (we discuss indicators in greater detail later in this chapter).

In order to provide further illustration of the differences between goals and objectives, figure 3.1 shows three examples for a hypothetical program. Notice that in all three examples the goals are broadly defined in terms of what the planner wants to change with the implementation of the program—for example, the program should, broadly speaking, increase financial inclusion. That is one of the reasons the program is being conducted. But such a goal tells us nothing about how to actually get such inclusion and how we would measure it. That is the role of an objective. Here, the objective is to "Raise the percentage of households with bank accounts from 5 percent to 10 percent over a period of one year following a financial education intervention."

Notice the specificity of the objective; it tells us how we would operationalize the goal of "inclusion" (through changes in the number of households with bank accounts) and provides both the magnitude of change and the expected time period in which the change is expected to occur. Such magnitudes need to be realistic, of course; in this case, one would surmise that the planner relied on what was found in the situation analysis that indicated an increase of 5 percentage points is a realistic

FIGURE 3.1 EXAMPLES OF GOALS VERSUS OBJECTIVES

GOAL: Increasing financial inclusion	**OBJECTIVE: Raise the percentage of households with bank accounts from 5% to 10% over a period of 1 year following a financial education intervention**
GOAL: Improving household money management among women	**OBJECTIVE: Reduce the percentage of women in the target population reporting "problems with day-to-day money management" from 60% to 20% over a period of 1 year**
GOAL: Increasing formal savings rates among the poor	**OBJECTIVE: Raise the percentage of individuals saving in a formal account from 20% to 25%**

expectation for this target population in the one-year time period. Doing so ensures that the results of the impact evaluation surrounding this goal and objective are ultimately meaningful, because there is a context for interpretation underlying them.

The process of establishing goals and objectives clearly makes sense when the evaluation is prospective. But is it also important in retrospective evaluations? Surprisingly, it is. Those conducting such evaluations after the fact should not be surprised if the goals and objectives of programs that have been operating for years are still unclear, even to the program's own staff members, or if they have changed over time. Alternatively, goals may have been set but the objectives may not have been sharply stated.

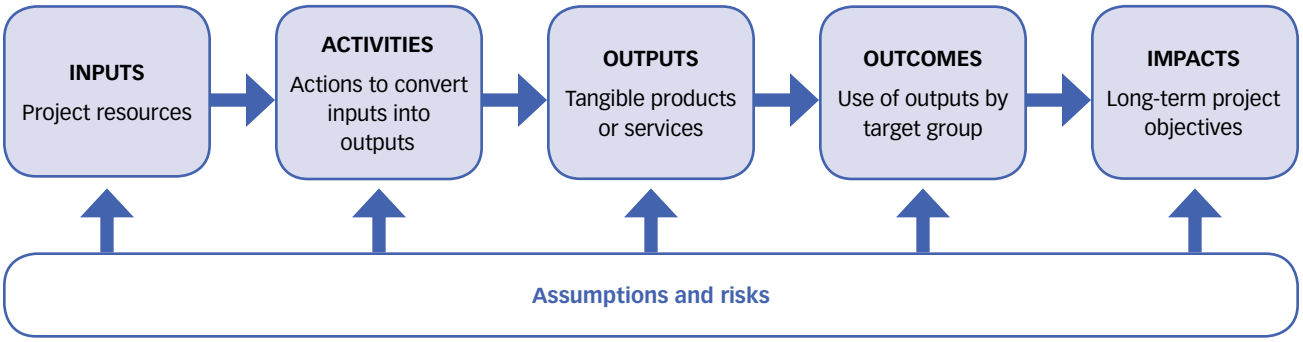
3.2 HOW CAN EVALUATORS CAPTURE A PROGRAM'S THEORY OF CHANGE?

The link between a program's design and its goals can be understood through the program's **theory of change**. Every financial capability program should be based on an underlying theory of change—which simply refers to a logical, sequential argument for how and why an intervention will deliver the desired results, together with any assumptions necessary at each step of the way. The program's theory of change should be set down explicitly at program inception—following the thorough situation analysis discussed above—and used to guide evaluation questions. (In a retrospective evaluation, however, evaluators may have to try to develop the theory of change after the fact.)

How does one depict a program's theory of change in a way that is useful for program evaluation? There are a lot of tools and frameworks that can be used for such an exercise, but we find that something called a **results framework model** is one of the simplest and most useful alternatives to depict the various stages in the program's theory of change, from inputs to final outcomes. Figure 3.2 shows an illustration of a results framework model, explicitly mapping the planned program logic in terms of the program's expected inputs and outputs, as well as key assumptions, preconditions, and risks that could affect the project outcome.

A program's theory of change should be set down explicitly at program inception and used to guide evaluation questions.

FIGURE 3.2 AN EXAMPLE OF A RESULTS FRAMEWORK MODEL



The basic components of all results frameworks are shown in the figure in the order they are shown. Below we describe each term, and use illustrative examples from the RTF pilot program in South Africa. Recall that this financial capability program uses a social marketing instrument to help increase residents' financial capability in terms of managing household debt; the program is explicitly seeking to change behavior in the approaches to debt resolution by over-indebted households.

As part of that effort, program designers began by building a results framework to lay out the objectives, activities, and related indicators of the financial education storyline in *Scandal!*.

- **Inputs** refer to all resources needed to carry out program activities, such as staff, materials, and finances. In our South Africa example, inputs include the costs of creating the show, including production, salaries for actors and writers, and so forth.
- **Activities** include all the actions and tasks that convert the inputs into outputs. These often resemble tasks on a to-do list or include work performed. For the example program above, activities include writing, producing, filming, and distributing the show.

Outputs are often confused with outcomes, which follow after. Outputs are under the program's control but outcomes may not fully be.

- **Outputs** are the products and services that the project produces through its activities. We can think of these as the program's deliverables. Outputs are often confused with outcomes, which follow after, but they are very different. Unlike outcomes, outputs don't tell us whether the program (in our case, the financial education storyline in *Scandal!*) had an impact on those trained. In the case of *Scandal!*, the outputs are the episodes shown on TV.
- **Outcomes** are the results or changes in the target population that occur from using the outputs (products and services). They are often short-term or intermediate changes and, unlike outputs, they may not be fully under the program's control. In other words, a program can provide a product or service (output), but it can't guarantee that such a product or service will change behavior (outcome). In the example program, outcomes would be improved understanding of household finances and perception of the importance of managing debt.
- **Impacts (**sometimes referred to as **final outcomes)** are overall program goals, such as changes in behavior or welfare of the target group, which come about as a result of the project outputs but, as with short-term outcomes, may also be affected by other, external factors. Final outcomes are often achieved only over a long period of time. In the context of the example we have been using here, an example of a final outcome could be the reduction of total household debt.
- **Assumptions** are expectations about the program components and implementation and about outside factors that can affect program operation or success. Assumptions should reflect the formative research discussed previously in determining project strategy, goals, and objectives. In the example we are using, a key assumption could be that people who have significant household debt will watch the soap opera. If that assumption is wrong and the program therefore does not reach the targeted participants, there is a risk that the recruiting activity will be ineffective, which would then likely threaten the program's success.

For some programs, intermediate outcomes may be expressed as changes in knowledge and attitudes, while final outcomes may be expressed as behavioral changes.

The distinction between outcomes and impacts (or alternatively, intermediate versus final outcomes) depends on the nature of the financial capability program and the scope of the objectives that it sets for itself. For instance, the final outcome of a more modest information-dissemination financial capability program may be to raise awareness (a change in knowledge or attitudes); as such, it may have no intermediate outcomes. However, by definition, financial capability programs are concerned with behavioral change, whether as an intermediate or final outcome. For some programs, intermediate outcomes may be expressed as changes in knowledge and attitudes, while final outcomes may be expressed as behavioral changes. Large-

scale financial capability programs may define intermediate outcomes as changes in knowledge, attitudes, and behavior, while aiming for final impacts such as changes in financial status.

In designing a program from scratch, it makes the most sense for evaluators to start with objectives and then work backward to the input requirements (iterating the process as needed). However, for an existing program where evaluators are evaluating retrospectively, it is often easiest to begin by making a list of all program activities and progressing from right to left, linking the inputs to the objectives. As with the need to sometimes define goals and objectives—even in established programs—the same is true about the results framework, which also should never be taken for granted. Often, evaluators find that program managers and stakeholders who may have been immersed in the day-to-day activities of the program for years disagree on important elements of the program.

However, regardless of whether evaluators are working prospectively or retrospectively, working through a results framework is critical because it clearly articulates a consensual commitment among stakeholders to the program's theory of change, clarifies uniform terminology, and provides a clear foundation for evaluation measures that tie together the program design and the M&E framework.

3.3 HOW CAN EVALUATORS DEVELOP EFFECTIVE INDICATORS?

Going down deeper into the development of a robust M&E system, having a set of high-quality **indicators** is crucial—especially given logistical, monetary, and resource constraints. What is an indicator? It is a quantitative or qualitative variable that measures achievement or progress toward a specific goal or objective. Once the program components in the results framework are mapped—the inputs, activities, outputs, and outcomes—indicators should be determined for each component from inputs to outcomes.

What constitutes good indicators in the results framework? The following tables capture the relationships between the stated objectives, the activities that are being implemented to accomplish those objectives, and the indicators that can measure whether the activities accomplish the objectives.

The objectives are divided between one focused on improving knowledge and one focused on improving attitudes and behavior, with each having corresponding activities to be carried out. Indicators for *Scandal!* were informed by consulting with researchers from the Financial Education Fund. The initial list of indicators included a large number of possible indicators for the different dimensions of financial capability

BE SMART WHEN CONSIDERING INDICATORS

A frequently used and helpful description of high-quality indicators is SMART:

Specific: Clearly and precisely defined.

Measurable

Attributable: Clearly linked to the objective, outcome, or output being measured.

Realistic: Simple and easy to interpret, feasible to gather.

Targeted: Variable, between and within subjects over time.

TABLE 3.1 EXPECTED OUTCOMES, ACTIVITIES, AND INDICATORS IN THE *SCANDAL!* PROGRAM

EXPECTED OUTCOMES/IMPACTS	ACTIVITIES	INDICATORS
1. KNOWLEDGE Increase the knowledge and understanding of low-income South Africans about personal financial management, with a particular focus on loans	▪ Include relevant financial-capability–building messaging in *Scandal!* storyline ▪ Specific topic areas will all be related to debt and will include but not be limited to: – Sound financial management (using credit wisely, budgeting, and setting goals) – Getting into debt (impulse buying, living beyond one's means) – Getting out of debt (seeking debt counseling, assessment tools, debt recovery)	▪ Measure changes in knowledge about types of financial services, specifically: – Loans (e.g., definition of a loan, advantages and disadvantages of credit) – Borrowing sources and options – Elements of loans (e.g., interest rates, loan terms, fees, penalties, delinquency policies) – Differences between formal and informal products – Rights of client and recourse options – Strategies for managing and reducing debt
2. ATTITUDES Improve financial capability attitudes among South African consumers related to debt management and help-seeking	▪ Develop messaging about the change of attitudes and include in *Scandal!*	▪ Measure: – Changes in confidence to ask relevant questions and negotiate terms – Ability to critically evaluate credit providers – Caution in borrowing – Strength to say no to unfavorable terms – Confidence to present complaints – Discipline to follow a debt-management plan
3. BEHAVIOR Improve financial capability behaviors among South African consumers related to debt management and help-seeking	▪ Develop public service announcement/call to action to consult NDMA and include in *Scandal!* storyline ▪ Develop message on possibility to request a personal credit report from the credit bureaus and include in *Scandal!* storyline	▪ Measure number of consultations to the NDMA website and number of calls to the hotline ▪ Measure numbers of viewers requesting a credit report ▪ Measure changes in maintaining an emergency savings account, making a plan to reduce debt, making loan payments on time, maintain an acceptable debt-to-income ratio

competencies and not all of them were included in the evaluation; the evaluation team selected only the most relevant ones.

Note that each input, activity, output, and intermediate- and long-term outcome has an indicator to concretely measure it. In an actual program, it is certainly possible—even likely—that multiple indicators could exist for any specific input, activity, output, or outcome.

In selecting indicators, evaluators should bear in mind some common practical decisions (and implicit trade-offs). Some common issues related to these arise in the case of financial capability programs.

TABLE 3.2 EXPECTED OUTCOMES, ACTIVITIES, AND INDICATORS IN THE FEATURE FILM PROGRAM IN NIGERIA

EXPECTED OUTCOMES/IMPACTS	ACTIVITIES	INDICATORS
1. KNOWLEDGE Provide individuals with knowledge about financial management and planning with a focus on savings	▪ Create a movie dealing with financial management and planning themes – *The Story of Gold* is a movie being produced by Nollywood and distributed by Credit Awareness – It tells the story of identical twin sisters growing up in Nigeria, who, although identical in appearance, make different decisions when faced with different financial choices that affect their lives and those around them and ultimately lead them down different paths. – The movie aims to teach low-income individuals with limited formal education some of the core concepts around financial planning under the motto of "Safe savings and responsible borrowing"	▪ Measure understanding of purpose of saving and pros and cons of saving with different institutions ▪ Measure knowledge about what savings instruments are available ▪ Measure understanding of the implications of being in debt
2. SKILLS Equip people with the necessary skills to be financially capable and act in a financially responsible manner when looking to save and budget		▪ Measure understanding of what steps need to be taken to open a bank account and when to save and when to use savings ▪ Measure ability to construct a budget
3. ATTITUDES Increase confidence and trust in formal savings institutions; improve perceptions about access to institutions for low-income or otherwise marginalized groups		▪ Measure perceptions of barriers to entry to financial institutions ▪ Measure motivation to engage with financial institutions ▪ Measure reported intentions and plans to start saving and budgeting for future investments
4. BEHAVIOR Improve individuals' behavior related to secure savings activities		▪ Measure number of new savings accounts (or applications for savings accounts) ▪ Measure savings rate
5. BUSINESS PERFORMANCE (longer-term impacts) Help make microenterprises more financially sustainable, efficient, and successful		▪ Measure turnover and profits, cash flow, and lending terms

One issue has to do with the distinction between direct and indirect indicators. A **direct** indicator refers directly to the subject of interest or the target population and is often the most preferred choice in terms of specificity. An **indirect** indicator is a measure that approximates or represents a phenomenon when a direct measure is not available. All things being equal, direct indicators make the most sense. But stakeholders might choose to use indirect indicators in situations where direct indicators cannot be measured, such as in the case of fundamentally intangible variables like feelings of well-being; they may also choose to go with indirect indicators if what you need to measure is considered sensitive or difficult to capture for other reasons, for instance, wages earned. One might also choose indirect indicators if they are more cost-effective and feasible. For example, evaluators may prefer to take a direct approach in creating financial knowledge outcome indicators, basing those

A direct indicator refers directly to the subject of interest or the target population; an indirect indicator is a measure that approximates or represents a phenomenon when a direct measure is not available.

indicators on performance in knowledge tests. However, doing so may be either too expensive or simply not feasible given the program's limitations. As such, evaluators might instead use an indirect approach to measuring knowledge, basing indicators on self-reports of perceived financial knowledge or confidence in financial ability. While this approach provides useful results, it is important to realize its limitations. Consumers often think that they know more than they actually do—a common finding that has been demonstrated not just in financial matters, but also across a wide range of knowledge and abilities. Whereas actual and perceived knowledge are often correlated, this correlation is often moderate at best. Hence, while circumstances may not always permit an objective assessment, caution should be taken when using perceived knowledge as a simple proxy for actual knowledge. Similarly, self-reported behavior may not always correspond to measured behavior.

Another issue in selecting indicators relates to the use of **compound** or **composite indicators**, which combine different components into a single measure. For instance, evaluators may decide to combine scores on a series of questions on financial knowledge and behavior into a single index of financial capability. Compound indicators provide more information, and fewer may be needed to capture the entire program; however, such indicators are often difficult for evaluators to "conceptually unpack," often requiring evaluators to make some form of judgment call to do so. In financial capability programs, in particular, evaluators want to avoid conflating knowledge, ability, and behavioral intent with actual behavior as much as they can when they develop indicators. While financial knowledge may be related to financial skills (e.g., negotiating mortgage terms, navigating an investment website) and behavioral intent (e.g., mutual fund fee minimization), financial knowledge does not necessarily imply financial capability, and vice versa. Skills and actual behavior are also likely to be influenced by other factors, such as access to resources, social networks, etc. Such other factors must be accounted for if evaluators are going to truly gauge how much changes in knowledge actually led to changes in behavior. Thus, the distinctions among actual knowledge, perceptions of knowledge, the ability to use that knowledge, and actual behavior are nontrivial.

Compound indicators are often difficult for evaluators to "conceptually unpack," often requiring evaluators to make some form of judgment call.

3.4 USING EXISTING RESOURCES FOR INDICATORS

Finally, many evaluators need to decide whether to use "standard" indicators (such as the examples given in table 3.1) versus indicators customized for the specific program. Using standard indicators that are common to other evaluations helps ensure comparability across programs, but doing so could mean that the indicators may not be as tightly connected to program goals as one would want. In practice, program indicators should be driven by program components, not vice versa.

Still, there is often no need to reinvent the wheel: standard indicators are often a good starting place for developing more customized ones, as long as evaluators avoid using standardized indicators when their content does not match that of the program. This is especially true for outcome indicators. Below, we describe two important and highly relevant new resources from the World Bank that evaluators can draw upon for their own work.

Evaluators can use standard indicators as a starting guide for developing more customized ones.

3.4.1 Measuring financial capability: the World Bank financial capability survey

While a host of survey instruments have been used to assess financial literacy and capability in other settings, the financial capability measurement work funded by the RTF and managed by the World Bank developed new methods for assessing levels of financial capability in a way that is relevant for low- and middle-income environments and that can be used consistently across countries to conduct international comparisons. The measurement instrument developed by the RTF provides a diagnostic tool that evaluators can use to identify the key areas financial capability (behavior, skills, attitudes, and knowledge) and which of these areas need improvement. The survey instrument can also be used to inform the design of targeted interventions in the area of financial education and financial capability enhancement, by helping identify especially vulnerable groups in terms of financial capability. In order to define an appropriate conceptual framework and to develop a survey instrument that is suitable for use across different countries and income levels, the World Bank used a rigorous instrument-design process involving both qualitative and quantitative research methods to understand how the concept of financial capability can be measured in these settings. For this purpose, the RTF has involved an international team of experts in financial capability and in questionnaire design, and it is supporting research projects in a group of 11 low- and middle-income countries (Armenia, Colombia, Lebanon, Malawi, Mexico, Namibia, Papua New Guinea, Tanzania, Turkey, Uruguay, and Zambia) to help develop and test the methodology.

Appendix C (available in the online version of this report and accessible at www.finlitedu.org) includes the report by the World Bank RTF Financial Capability Measurement Program team describing the development of its Financial Capability Survey. The survey is reproduced in whole in appendix E, which is also available in the online version of this report and accessible at www.finlitedu.org.

3.4.2 Measuring financial inclusion: the Global Findex questionnaire

An important category of outcome indicators is financial inclusion. While systematic indicators of the use of different financial services had been lacking for most econo-

mies, particularly low- and middle-income environments, the Global Financial Inclusion (Global Findex) database, released in April 2012 provides such indicators. This short list of key indicators covers a range of financial behaviors on saving, borrowing, making payments, and managing risk. In addition to providing a reference for question format and language, the indicators developed by the World Bank team were fielded on more than 150,000 nationally representative and randomly selected adults age 15 and above in those 148 economies, providing a natural and accessible way to benchmark the characteristics of a program's population.

The Global Findex questionnaire is reproduced in whole in English in appendix E of the online version of this report (accessible at www.finlitedu.org) but is also available in 14 other languages on the Global Findex website (http://www.worldbank.org/globalfindex).

KEY POINTS

Having a clear map of each program component and the required indicators is a useful practical basis for generating an overall M&E framework that systematically captures program components and the types of information that evaluators would in theory need to be able to assess each component. In the next chapter, we discuss using the components laid out in this chapter to develop and implement a monitoring plan for financial capability programs.

FURTHER READING

Organisation for Economic Co-operation and Development. 2005. "Improving Financial Literacy: Analysis of Issues and Policies." As of October 19, 2012: http://www.oecd.org/daf/financialmarketsinsuranceandpensions/financialeducation/improvingfinancialliteracyanalysisofissuesandpolicies.htm.

Alba, J. W., and Hutchinson, J. W. 2000. "Knowledge Calibration: What Consumers Know and What They Think They Know," *Journal of Consumer Research* 27 (2): 123–56.

Demirguc-Kunt, A., and L. Klapper. 2012. "Measuring Financial Inclusion The Global Findex Database," The World Bank Development Research Group Finance and Private Sector Development Team. As of October 19, 2012: http://www-wds.worldbank.org/servlet/WDSContentServer/WDSP/IB/2012/04/19/000158349_20120419083611/Rendered/PDF/WPS6025.pdf.

Lichtenstein, S., Fischhoff, B., and Phillips, L. 1982. Calibration of probabilities: The state of the art to 1980. In: D. Kahneman, P. Slovic and A. Tversky (eds.), *Judgment Under Uncertainty: Heuristics and Biases*, Cambridge and New York: Cambridge University Press.

Yates, J. F. 1990. *Judgment and Decision Making*, Englewood Cliffs, NJ: Prentice Hall.

PART II

Conducting M&E for financial capability programs

CHAPTER 4

Monitoring

As we discussed in chapter 2, monitoring and evaluation (or M&E) go hand in hand. Monitoring is critical, both in and of itself but also as the underpinning of evaluations. Monitoring a financial capability program from its inception helps ensure that those who manage it—along with policy makers and other stakeholders—can see that work plans and milestones are going as planned in terms of how that program is delivered and, if not, that plans are in place to rectify any problems. And, as noted in chapter 2, such information and data can be fed into the planned evaluations as well.

Monitoring is critical in and of itself and as the underpinning for other evaluations.

Chapter 2 defined monitoring briefly. In this chapter, we dig a little deeper and focus on the key components of monitoring systems, starting with a brief reiteration of what monitoring is and why it is important and then moving on to developing and implementing a monitoring plan, bring the monitoring information together, and using the results of monitoring in program management.

4.1 WHAT IS MONITORING AND WHY SHOULD WE DO IT?

As noted earlier, monitoring involves regularly collecting, analyzing, and reviewing information about how a financial capability program is operating. It is a continuous and systematic process that can be used to gain rapid, real-time insights into the program as it develops and on the progress it is making toward meeting work plans and milestones. It can also be used to track the use of the program's budget and to identify any emerging threats to its success.

Because it is a systematic process that continuously collects and analyzes information on project implementation and delivery against planned results, it usually makes sense that monitoring is grounded in the results framework that the program sets up and that we discussed in chapter 2 and in more detail in chapter 3, involving inputs, activities, outputs, and (in some cases) outcomes. Because monitoring supports day-to-day program management, it is an everyday activity and conducted over the full program cycle.

Because monitoring supports day-to-day program management, it is an everyday activity and conducted over the full program cycle.

Why should a monitoring system be established? There are many reasons, but most directly, monitoring can provide managers, staff, and other stakeholders with routine indicators on the implementation and progress of the program, as well as facilitate timely adjustments to emerging threats and opportunities. If, for example, a financial capability program requires that specific activities occur at specific times, a well-laid-out monitoring system will tell program managers whether those activities have occurred when they were supposed to and, if not, will aid in making sure that they do down the road; in a sense, monitoring can serve as a sort of early warning system.

Monitoring can serve as a sort of early warning system.

Another reason to institute a well-planned monitoring system is to generate information and indicators for use in process evaluations and impact evaluations. Monitoring systems and process evaluations are particularly closely related and often share similar infrastructure and data sources. Both activities are important, because both help improve the efficiency and effectiveness of a program, in addition to increasing accountability with stakeholders. A rigorous and explicit monitoring plan is critical for systematic management of project resources and, importantly, can help create a common understanding of the project among stakeholders and funders. A robust process evaluation—with support from a thorough monitoring system—can help develop responses to unforeseen challenges and weaknesses of a program, as well as bolster its strengths.

And, as discussed later, monitoring and process evaluations are important complements to an impact evaluation. These benefits make it worthwhile to invest in developing an integrated M&E framework at an early stage of the project, and to continuously refine and update the system throughout the program cycle.

4.2 HOW CAN WE DEVELOP AND IMPLEMENT A MONITORING PLAN?

As a first step, an effective monitoring system requires a plan that describes how the system will be implemented. A good start is to build off the results framework described in chapter 3. This framework can serve as the basis for a more detailed plan for implementing the monitoring system.

Table 4.1 shows what such a plan could look like for an example loosely based on a Financial Education Fund (FEF) project in Northern Ghana that aims to increase financial literacy of farmers through capacity-building activities, either through group-based workshops or group-based workshops alongside one-on-one training. The

TABLE 4.1 EXAMPLE MONITORING IMPLEMENTATION PLAN

CATEGORY	INDICATORS	FREQUENCY	INFORMATION SOURCE	ASSUMPTIONS & RISKS	STAFF
Inputs					
Training material printed	Number of materials printed	Monthly	Site reports		Site managers
Activities					
Objective: Training workshops are conducted	Number of workshops organized at each site	Monthly	Progress report	Workshop facilities are available as planned throughout project period	Survey supervisors
Outputs					
Objective: Farmers participate in financial literacy course	Number of individuals completing training course	Monthly	Training participant sign-in sheets	Individuals enrolled on phone will attend on-site training	Site managers; program manager
Outcomes					
Objective: Participants have better knowledge of their financial situation	Percentage of farmers who know their precise personal savings balance	Baseline and end of year 1	Questions in baseline and end-of-project surveys, and bank records	Participants will be willing to tell us their balance in the survey, and allow us to examine their bank records.	Program manager
Final outcomes					
Objective: Participants increase their net savings	Percentage increase in number of farmers with an increase in net savings after 12 months	End year 1	End-of-project survey		Program manager

table is oriented around the results framework—a useful way to structure the plan. Here we use the delivery of the workshops to illustrate what a monitoring implementation plan would look like. The table is driven by the results framework categories; within each category, we have the specific objective followed by the desired indicator and the frequency of reporting, as well as the specific information source to be used, any pertinent assumptions and risks, and the staff who will be responsible for this monitoring activity.

A monitoring implementation plan should be built around a results framework that includes inputs, activities, outputs, and (often) outcomes.

For example, to deliver the workshops, we list the training materials as inputs, which are obviously critical. One key indicator that we want to monitor relative to these materials is how many of them were actually printed; thus, in the indicator column, we list the indicator we want to measure: the total number of printed materials. We also want to know how frequently to monitor the indicator; the frequency here

is monthly, because we want to be able to react quickly if we run out of materials and, thus, need to have current information. The logical source of this information is the reports from the sites where the training materials are prepared. As for who is responsible for collecting the information, it makes sense that on-site managers, who will have access to that information continuously, be the ones responsible.

In this example, the "inputs" category does not have a set of assumptions or risks associated with it, but the "activities" one does: It assumes that the facilities where the workshops will be held are available in the project period. Making such an assumption explicit in the plan ensures that the risk of not having access to a meeting room is understood, because if the assumption does not pan out, it will affect the number of training workshops the program can conduct.

When devising such a plan, it is important to consider carefully the type of information needed, the choice of the information source, and the frequency of data collection. In practice, for many programs, local-level units will compile data on a monthly or quarterly basis, which is the case for all but the "outcomes" category in the example above. But thinking through how frequently data are needed is not a trivial issue; in principle, having more and more frequent reports could be valuable, but doing so in practice will likely pose serious issues for those responsible for collecting all the data. Thus, the frequency of data collection should be balanced against the capacity of the designated staff to carry it out (including literacy/numeracy skills and the ability to schedule recordkeeping as part of already preexisting regular program duties). It also should be balanced against the program's ability to actually respond to the information. We discuss this point a little later on in this chapter, but it is worth stressing here that those making use of the data must be able make use of it; if data are collected with too much frequency, it may not get used in a timely manner, thus overwhelming all parties and defeating the purpose of collecting it so frequently in the first place.

When devising an implementation plan, carefully consider the type of information needed, its source, and how frequently such information must be collected.

For the project in the above example, the project team defined a reporting schedule and a protocol reference sheet for each indicator. This sheet lists the name, objective, and definition of the indicator alongside the unit of measurement, data source, and how it will be analyzed. Explicit codification of these details helps all team members to understand how the monitoring will be done. An example protocol reference sheet drawn from the FEF project in Northern Ghana is shown in figure 4.1. In this example, the protocol reference sheet is filled out for an indicator defined as the percentage of farmers who know their personal savings account balance.

Once the monitoring plan is set, it should be reviewed in its entirety to ensure that the burden of responsibility is fair and reasonable and that accountability is not overly concentrated. As noted in the example above, it is fair and reasonable (indeed, makes the most sense) for site managers to collect data to measure the indicator on

FIGURE 4.1 EXAMPLE INDICATOR PROTOCOL REFERENCE SHEET

Name of Indicator: *Insert the complete name of indicator*
% of farmers who know precisely their personal savings balance
Objective to which Indicator Responds: *Why are you measuring this indicator?*
The indicator responds to objective 2: To increase the use of financial services mainly through savings and cash-crop loans.
For those farmers who operate savings accounts, the indicator will seek to assess the number (%) who know their savings balance.
Definition of the Indicator: *Unpack as much as possible the specific definition of this indicator. When will the individual/Unit be counted as reached?*
The number (%) of farmers who know how much money is left in their savings accounts.
Unit of Measurement and Disaggregation: *In what unit will this indicator be captured and is there any disaggregation (male/female, age, etc.)*
Number of farmers, disaggregated by gender (male/female); Groups (A, B1, B2, and C) during the end-of-project evaluation; Years of participation in Primary Farmer Based Organization; and Literacy.
DATA SOURCE
Data Source: *Where will the data be collected?*
Data will be elicited from farmers, Association of Church-Based Development Non-Governmental Organization Financial Education Project and banks through questionnaires during the baseline study and end-of-project evaluation.
DATA ANALYSIS
Data Analysis: *How will the data from this indicator be analyzed? Who will be responsible for analysis? How often are the data analyzed? Specify when the analysis will be conducted.*
The data will be collected during the baseline study (August–September 2010) and end-of-project evaluation stage (June–August 2011), and analyzed during those periods using the SPSS statistical package. These analyses will be done by external consultants.

how many printed training materials are available at each workshop, because they are actually there. You will need to work with designated staff to ensure that they understand their roles within the overall system. You should discuss with them their specific reporting responsibilities, including the frequency and sources of information to be used. It is important that the monitoring tasks are feasible at all levels. Program staff will also be more motivated to collect and report monitoring data when they understand the purpose of the monitoring system and how it can benefit the program and possibly their own work. It is important to ensure that the monitoring process does not become so burdensome as to defeat its purpose.

Monitoring plans should be reviewed to ensure the burden of responsibility is fair and accountability is not overly concentrated.

While having an organizing matrix such as that outlined in the above example makes sense for monitoring all programs, those implementing more complex interventions may want to complement the matrix with other project management tools that provide a closer look at particular aspects of the monitoring process; such tools can also help to operationalize the matrix for day-to-day operations. For example, there may be sequences of processes where one process must take place before another one does; it may, for example, rely on the data from that process. Flow diagrams could be used in this case to clarify the sequence of processes. And recordkeeping

BOX 4.1 MONITORING IN PRACTICE: AN EXAMPLE FROM THE RTF PILOT PROGRAM IN MALAWI

In some programs it is possible to collect meaningful monitoring data automatically. For instance, an RTF pilot program in Malawi aims to encourage formal savings by agricultural workers. Rather than paying workers in cash, workers' wages were directly deposited into bank accounts. The bank then records activities on these accounts, such as the time and amount of deposits and withdrawals. With proper consent from the account holders, these records can be used to monitor the program's progress and to adjust the program operations if necessary. And, of course, this highly frequent and detailed administrative data can also feed into the process and impact evaluations.

charts can be used to document specific staff responsibilities for performing and supervising monitoring activities.

While financial monitoring is not typically part of M&E plans, it is critical for both program management and program accountability purposes.

Before leaving this section, it is worth noting that financial monitoring is a critical component of M&E plans but is often not part of them. The reason for this is that program finances are often managed separately from actual program implementation; as such, financial monitoring is typically conducted by a program accountant rather than a program manager and is tied to the budget cycle.

Still, while costs may not be physically or functionally part of an M&E system, the costs of inputs and activities are important monitoring indicators, and tracking expenses is critical for project management. In the example above, the cost of the training materials—in terms of designing and printing them—that serve as inputs to the training sessions must be accounted for. Good bookkeeping facilitates accountability within the program and is often required by external supporters or funders. But beyond accountability purposes, having reliable cost data will serve as a key input in any cost-effectiveness and cost-benefit analyses. We discuss this more in chapter 11, when we talk about such analyses in more detail.

4.3 HOW DO WE BRING MONITORING INFORMATION TOGETHER?

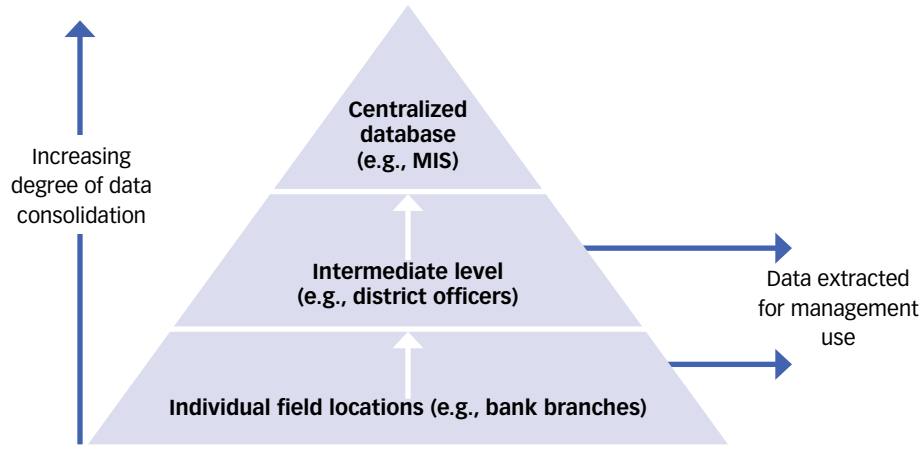
Not surprisingly, the output of the monitoring plan is a large amount of data and information. As the example in table 4.1 shows for the training session part of the FEF financial capability program in Northern Ghana, large amounts of data are being collected on a frequent basis.

Figure 4.2 shows that data and information typically flow from the bottom up, where they get consolidated. For example, for large programs, information is collected

and consolidated at individual field locations (e.g., individual bank branches, local community centers). The data are then examined, collated, and sent on to an intermediate level (e.g., a district office), and from there sent on to create a centralized database. At each administrative level, some relevant information can be extracted for management use.

At each stage of the process, it is important to balance what information is collected and retained and what information is sent on and by whom. Of course, some

FIGURE 4.2 PYRAMID OF MONITORING INFORMATION FLOWS



amount of duplication across individuals or program administrative levels is to be expected and is, in fact, important for recordkeeping and cross-checking. But it is still important to maintain a balance, such that only the most necessary information is aggregated and sent upwards, especially when the next level consolidates a vast number of lower-level units. Focusing only those data that are critical for assessing the daily program operations will reduce the burden on the field staff to collect the data and ensure that the information is accessible to managers.

Having both clear protocols and forms for data entry and reporting is one way to ensure the right data in the right form get to those who need to use them.

As part of this process, it is important to have both clear protocols and forms for data entry and reporting. In the first case, having set guidelines and schedules for entering and/or processing this information regularly at the field level can greatly increase the quality and efficiency of data collection. At each stage, quality checks and validations should be conducted, either by examining the aggregate data, conducting random audits, or both.

As for the forms, collection and reporting forms that simplify and standardize information can facilitate high-quality data entry and review. In the above-mentioned FEF

project in Northern Ghana, the project team designed a simple sign-up sheet for participants who ask for the training name and location, in addition to recording the names and contact details of the participants. Providing these sheets to each location reduces the chance that the participant list is incomplete or that some critical information about the training (such as the location) is missing.

Note that if data are being collected on forms that are being newly developed as part of the monitoring system, it is important to design and test the forms with a view toward their use in the field. If those charged with actually entering the data do not understand the form or how to use it, then the data collected will not be useful. Another good reason for field-testing the forms is that changing paperwork **after the fact** can be highly disruptive to field staff; thus, it is wise to ensure from the start that forms are as complete as possible (including identifiers for follow-up), but not overly burdensome.

Typically, as shown at the top of figure 4.2, the highest degree of consolidation relies on a centralized, computerized **management information system (MIS)**. Computerization has significant advantages, including the ability to validate data as it is entered, perform automatic calculations, consolidate information quickly, and generate backups. Not to mention that the use of a computerized system often helps in transferring the data and in making it more likely that monitoring data are easily accessible and usable in the context of a larger process or impact evaluation.

A highly developed MIS may be used from bottom up, including an interface that allows program data to be entered electronically by implementing staff or, in larger projects, by dedicated operators. Higher-level checks and data on outputs may need to be entered by supervisors. New approaches to computerized monitoring also make it possible to use mobile devices such as cellphones for data entry from the field.

Having said this, though, and in many cases relevant to this Toolkit, the use of manual data entry and management at lower levels may be more appropriate and quite sufficient, depending on the situation. There are many factors to consider when deciding between a computer or paper-based system, such as the:

- Nature of the program delivery, whether through a centralized point or by traveling field agents.
- Availability and maintenance needs of computer systems.
- Availability and training needs of staff with operating skill.
- Environmental conditions (e.g., heat, dust, humidity, reliability of power, and possibility of theft).

There is likely to be a trade-off between what many program managers may want and what implementing staff can or want to do. For example, program managers may prefer to receive reports in electronic spreadsheets, but implementing staff may find that entering data this way is cumbersome and distracts them from dealing with participants. In such cases, it may be better to track processes using paper and pencil and enter the data electronically at a later time and at a different level in the pyramid. Typically, data entry and review is done on paper up to the central level and then entered into a computer system.

When it comes to an MIS, there is likely to be a trade-off between what many program managers may want and what implementing staff can or want to do.

4.4 HOW DO WE USE MONITORING IN PROJECT MANAGEMENT?

A good monitoring system supported by a comprehensive MIS can guide project operations, track financial resources, provide managerial decision support, and facilitate communication with stakeholders.

In a common scenario for high-level communications and reporting to internal and external stakeholders, field reports are generally continuously produced, on a monthly or quarterly basis, with information at a centralized level produced biannually or annually—or in response to the program's administrative deadlines, requests for funding, and the like. Later parts of the Toolkit will expand on the use of such reporting overall.

However, apart from such high-level reporting, it is also desirable to build data utilization into the monitoring system itself, by explicitly incorporating a feedback process. In every period, after information is consolidated and reported, project implementers at each level can review the data to examine the aggregate period performance relative to previous periods. This ensures that the monitoring system supports all staff in their daily tasks and gives them opportunities to learn and respond directly to potential problems. The FEF project described above holds quarterly review meetings for this purpose.

In terms of the latter, because monitoring is ongoing and conducted in real time, a monitoring system can be used to identify potential problems as they arise. One way to do this is to include automatic "trigger points" tied to given indicators. When the specified indicators reach a specific level, certain responses are automatically initiated according to a predefined process in the monitoring plan. Trigger points can be negative and tied to "warnings," in which case the monitoring system can require an immediate remedial response by the management team, followed by frequent assessments to ensure the problem has been resolved. Action can then be taken and reported as part of the monitoring report itself.

TWO DIFFERENT USES OF TRIGGER POINTS

Warnings: notify managers when implementers have missed targets

Incentives: notify managers when implementers have met targets

Consider the example of the FEF training project from earlier. One of the key "activities" is the number of training workshops conducted at each site. The monitoring system can place a trigger or warning flag in place to notify program managers if the number drops, say, below four per month at a given site. Similarly, if the percentage of customers who complete the training course were to drop below the expected take-up rate, say, below 20 percent, then program managers would be immediately notified and could investigate potential causes and strategies to raise completion.

Of course, monitoring systems do not just need to register warnings; a "positive deviance" approach, for example, would link these trigger points to incentives. Monitoring systems can then be used to identify and deliver recognition to implementers who perform well and to give management the opportunity to identify and learn from best practices. This kind of monitoring can give credit to deserving program staff, build incentives to improve efficiency, build morale, and generate staff support for monitoring.

4.5 HOW CAN WE BUILD A ROBUST MONITORING SYSTEM?

The guidance provided in this chapter explains in broad terms how monitoring is conducted and used, but the reality is that any monitoring system for a specific program will need to be tailored to that program and to what makes sense for that program.

Having said that, there are some universal factors applicable to any program that will enable you to build a robust approach to monitoring. Table 4.2 provides a checklist of

TABLE 4.2 CHECKLIST FOR ENSURING A QUALITY MONITORING SYSTEM

Planning for monitoring	▪ Are the monitoring indicators clearly linked to the inputs, outputs, outcomes, and final outcomes they are meant to measure? ▪ Does the monitoring plan identify all the critical risks and assumptions? ▪ Are sufficient resources available for the ongoing monitoring activities, including staff and materials (such as reporting forms)? ▪ Are the program staff aware of their responsibilities in the monitoring plan?
Implementing the monitoring plan	▪ Are there instructions and established processes for collecting and reporting monitoring indicators and measures? ▪ Is the information and data needed for monitoring captured and reported in a timely and systematic fashion? ▪ Are data disaggregated at the levels required by the monitoring indicators? ▪ Is there a process for controlling the quality of the data at all levels? ▪ Are the necessary safeguards in place to ensure confidentiality and privacy of the program beneficiaries?
Using the results of the monitoring plan	▪ Are the monitoring plan and reports updated regularly?

these factors, grouped by planning for monitoring, implementing the monitoring plan, and using the results of monitoring.

KEY POINTS

The process of monitoring—regularly collecting, analyzing, and reviewing information about how a program is operating—is a critical part of any assessment of a financial capability program, both in and of itself and as an input to the process, impact, and cost evaluations discussed in the subsequent chapters.

A well-designed monitoring plan will help program implementers understand how a program is rolling out—whether the inputs and activities that are parts of the results framework are aligned with the program's objectives; and whether the delivery of those inputs and activities is proceeding as planned. Being able to see in real time any emerging problems allows implementers to rapidly adjust a program to make sure it gets back on course. A well-designed monitoring plan also helps ensure that critical stakeholders involved in the program's implementation stay abreast of the progress as the program unfolds.

FURTHER READING

General

IFAD (International Fund for Agricultural Development). 2002. "A Guide for Project M&E: Managing for Impact in Rural Development." Rome: IFAD. As of February 12, 2013: http://www.ifad.org/evaluation/guide/

Kusek, J. Z., and Rist, R. C. 2004. "Ten Steps to a Results-Based Monitoring and Evaluation System: A Handbook for Development Practitioners," vol. 289, World Bank Publications.

United National Development Programme (UNDP). 2002. *Handbook on Monitoring and Evaluating for Results*, New York: UNDP Evaluation Office.

Valadez, J., and Bamberger, M. (Eds.). 1994. "Monitoring and Evaluating Social Programs in Developing Countries," Washington, DC: The World Bank Economic Development Institute.

World Bank. 2004. "Monitoring and Evaluation: Some Tools, Methods and Approaches." As of February 12, 2013: http://lnweb90.worldbank.org/oed/oeddoclib.nsf/24cc3bb1f94ae11c85256808006a0046/a5efbb5d776b67d285256b1e0079c9a3/$FILE/MandE_tools_methods_approaches.pdf.

Technical

Davidson, E. J. 2004. "Evaluation Methodology Basics: The Nuts and Bolts of Sound Evaluation," Sage Publications, Inc.

CHAPTER 5

Process evaluation

As we discussed in chapter 2, process evaluations address a broad range of questions about the nature of a financial capability program's development and implementation, including its relevance, efficiency, and effectiveness. Such evaluations are used to help program managers understand how well a program is being implemented relative to the program's goals and objectives, how participants feel about the program and the services being delivered, what barriers implementers might be facing in implementing the program effectively, and how to address those barriers in real time to improve program implementation. The information from a process evaluation is also valuable to other stakeholders who have an interest in the program, such as funders and local authorities.

In this chapter, we discuss when it makes sense to conduct a process evaluation. Then, we lay out the steps that should logically be a part of process evaluations.

5.1 WHEN SHOULD WE CONDUCT A PROCESS EVALUATION?

The question of when, and whether, to conduct a process evaluation can be a difficult one to settle. When is a process evaluation good value for money? Do we need one if we already have a monitoring system in place? Should we spend our scarce resources on a process evaluation, or should we just aim for a bigger impact evaluation? Why should a monitoring system be established?

Part of the difficulty—which is reflected in the above questions—is that there is a great deal of overlap between what monitoring (and a monitoring system) does and what process evaluations do. Both focus on understanding the implementation of a financial capability program. As we mentioned in chapter 4, monitoring can provide managers, staff, and other stakeholders with routine indicators on the implementation and progress of the program, and can facilitate timely adjustments to accommodate emerging threats and opportunities.

There is a great deal of overlap between what monitoring and process evaluations do.

If, for example, the financial capability program requires specific activities to occur at specific times, a well-laid-out monitoring system will tell program managers

whether those activities occurred when they were supposed to and, if not, will aid in making sure that they do down the road. In a sense, monitoring serves—as we noted earlier—as a sort of early warning system.

There is no universally applicable answer to the question of whether one needs both a monitoring system and a process evaluation. It depends on the type of program, the resources available, and whether a process evaluation would actually help answer important questions about the intervention that a monitoring system would not. While monitoring systems can address whether inputs, activities, and services are being delivered as anticipated, process evaluations focus more on the effectiveness of those inputs, activities, and services relative to the program's goals and objectives.

Process evaluations are particularly important for new or complex interventions, such as those dealing with sensitive issues.

Above and beyond this distinction, there are some cases where it seems clear that process evaluations will make sense. Examples include financial capability programs that involve new or complex interventions, such as those involving multiple stakeholders or services, those dealing with sensitive issues, or those taking place in conflict areas. In those cases, process evaluations can be extremely valuable in helping practitioners identify emerging challenges and areas for improvement. For example, in the case of financial capability programs dealing with individuals' personal financial information (which can be considered sensitive), process evaluations can indicate whether the intervention's approach is deemed acceptable and safe to the intended beneficiaries.

5.2 HOW SHOULD PROCESS EVALUATIONS BE CONDUCTED?

Once a decision has been made to conduct a process evaluation, there are series of logical steps to take in the process, which are captured in figure 5.1.

FIGURE 5.1 STEPS IN CONDUCTING A PROCESS EVALUATION

5.2.1 Develop results framework

Although process evaluations are different from impact evaluations, they can both benefit greatly from using the results framework model discussed in chapter 3 to conceptualize the intervention and reflect the program's components (inputs, activities, outputs, and outcomes), what it is expected to achieve, and how it is expected to achieve it. In a process evaluation, this framework can inform the development of the key questions guiding the inquiry, something we discuss in the next step.

An important consideration in deciding on the focus of the process evaluation is whether the information collected will actually be valuable to the intervention, and whether the data will be used to make adjustments and improvements. It is easy to come up with a long list of interesting questions for a process evaluation; the challenge is to develop a process evaluation that will constitute value-added for the intervention.

It is easy to come up with interesting questions for a process evaluation; the challenge is to develop a process evaluation adds value to the intervention.

5.2.2 Develop process evaluation questions

A process evaluation can address many types of questions, but they are generally grouped into three main areas: **outputs**, **implementation/operations**, and **appropriateness/acceptability**. While the specific questions to be covered depend on the specific characteristics of the program and its context, figure 5.2 shows a sampling of common questions that broadly apply for many evaluators.

As noted above, some of the questions in a process evaluation, particularly those about program outputs, are also the focus of monitoring systems. However, monitoring is a **routine process of documenting or reviewing the implementation** while process evaluations provide an **overall assessment of implementation and delivery mechanisms**, at an early or intermediate point in a program's life, including whether improvements could be made. Process evaluations can be particularly useful in the pilot or early implementation stage of programs, where the focus is more on tweaking implementation to ensure a smooth operation rather than on the project's impact.

MONITORING VERSUS PROCESS EVALUATION

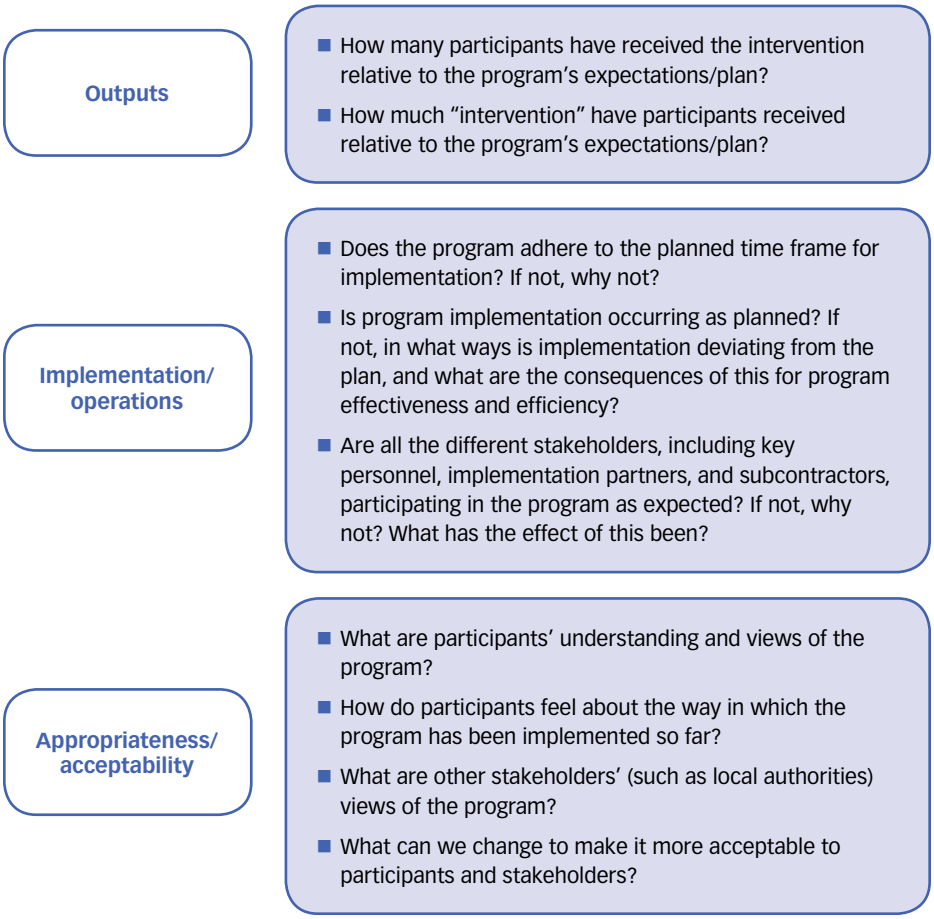
Monitoring: routine process of documenting or reviewing implementation

Process evaluation: overall assessment of implementation and delivery mechanisms

5.2.3 Collect necessary data

Once evaluators have settled on the main focus of the process evaluation and the types of questions they want to answer, data collection can begin. The data collected will of course depend on those questions, but here we talk about three types of data that can be used to inform a process evaluation: monitoring data, qualitative data, and quantitative data.

FIGURE 5.2 SAMPLE QUESTIONS FOR PROCESS EVALUATION



MONITORING DATA

In many cases, and where available, monitoring data can form the basis of process evaluation, especially when the evaluation's key questions focus on outputs, processes, and even finances. Monitoring data often includes details on how many beneficiaries a financial capability program is reaching, the demographic characteristics of those beneficiaries, and costs involved in implementing the program. For example, in the case of bank-based financial capability training, metrics could include the number of people attending workshops relative to both planned attendance and to all potential attendees, number of workshops provided, how much attrition there is from the workshops, how much the workshops cost, and so forth.

While monitoring tracks program management and delivery, process evaluation addresses larger issues of whether the program is performing as planned in terms of design and targeting as well as implementation. In this regard, monitoring can be viewed as an input to process evaluation. Indeed, process evaluations can and often do include an assessment of the monitoring system itself. While monitoring may be

an input to process evaluations, such evaluations can and do take place without a formal monitoring plan.

Process evaluations can range from very comprehensive to very narrow. As suggested in chapter 5, resources and time often dictate the choice on how to focus a process evaluation. For example, a comprehensive evaluation might choose to examine a number of questions about outputs, implementation, and acceptability of a program when a narrower study would only focus on a small number of issues in those areas.

Process evaluations can be comprehensive or narrow, depending on time and/or resources.

QUALITATIVE DATA

The data that come from monitoring systems are often quantitative data (e.g., the number of training workshops given), but process evaluators often use a range of qualitative methods to collect data. While a robust monitoring system may be able to provide quantitative answers to some of the implementation questions above, in many cases these answers may not be enough, especially when evaluators are interested in participants' experiences in receiving program services or implementers' experiences in delivering those services. Many of the questions in a process evaluation, such as those about the public acceptability of the intervention and participants' experiences, may best be answered through qualitative research, such as interviews and focus groups with stakeholders. Some process evaluations even rely exclusively on qualitative research.

BOX 5.1 RTF IN PRACTICE: PROCESS EVALUATION USING SURVEYS

In the course of the Russia Financial Literacy and Education Trust Fund (RTF) pilot program of a school-based financial capability program in Brazil, the evaluators conducted a survey of the school principals and teachers who were delivering the intervention. This survey was part of an assessment of how the program was unfolding. The data collection exercise also aimed to highlight areas for improvement in program delivery to inform the process of scaling up to the rest of the country. The surveys of teachers and principals happened at the end of the school year, at the same time as the student baseline and follow-up surveys. Because the researchers wanted to learn about program delivery, the data collection took place only in treatment schools.

The questions focused on whether the school received the necessary materials, how many classes received the intervention, how teachers used the materials to teach the program's content, and other issues about program delivery. These questions were administered by the researchers; an additional set of self-administered questions included the teachers' backgrounds, qualifications, and views of the financial capability content taught as part of the program.

The researchers plan on investigating the program delivery by looking at a range of indicators such as the percentage of schools where books were utilized and the percentage of teachers who thought books were very helpful. These statistics as well as a summary of qualitative data will also be included in the evaluation report.

Evaluators can use a range of qualitative methodologies to collect data in a process evaluation, including:

Use of protocols can help to ensure consistency in gathering data in interviews and focus groups.

- In-depth interviews
- Semi-structured and cognitive interviews
- Focus groups
- Desk review of documents and materials
- Mystery shopping and audit studies.

These methodologies and their implementation are described in more detail in chapter 5. As mentioned above, the data collection method used will depend primarily on the questions you need answered. For instance, the process evaluation of a school-based financial capability program may require observations of classes to assess how well the material is being delivered, as well as focus groups with students to explore whether the material is suitable given the intervention's aims.

QUANTITATIVE DATA

As noted, monitoring data are almost always quantitative in nature, but there are other kinds of quantitative data that can be used to good effect in process evaluations; such data can provide a useful way to quickly learn about the experience of larger groups. For example, a school-based financial literacy program may want to understand how teachers feel about the course material, how students respond to the lessons, and what practical challenges exist in different contexts (see box 5.1). Such information is not available in routine monitoring data, but it can be collected

BOX 5.2 RTF IN PRACTICE: PROCESS EVALUATION USING FOCUS GROUPS AND INTERVIEWS

One of the RTF pilot projects examines a financial literacy training program about the importance and benefits of saving in Bihar, India. In order to disentangle the mechanism through which attitudes and behaviors are shaped by the financial literacy intervention, the project team conducted a process evaluation using in person interviews and focus groups. These interviews were conducted with program staff and program participants immediately following the delivery of the intervention. The questions addressed a number of areas, such as service delivery (e.g., location and timing of training, feasibility of program approaches used), resources (e.g., were resources adequate, how to reduce resources without compromising program quality), service use (e.g., how were participants recruited, which participants took up the program), and program participant experiences (e.g., what was motivation for participating, opinions on length of training and content of the course).

quickly through a short quantitative survey. Although quantitative data generally provides less in-depth information than qualitative data, it often can be collected faster and can contain information about more people.

5.2.4 Analyze data

Once all the data collected for your process evaluation are in, you are faced with the task of figuring out what it all means for your program and how to put that information to best use. While there are no hard and fast rules about how to do this, it pays to be systematic in your analysis. It is important in this regard to remember that the data were not collected in a vacuum; they were collected to answer some specific process evaluation questions—questions that were, in turn, driven by the results framework that underpins the evaluation itself. Thus, it makes sense to go back to your process evaluation questions and identify the answers from among your data and ensure that the analysis of those answers is linked with goals and objectives that are drivers of the results framework.

5.2.5 Develop and implement solutions to problems and challenges

The final step in figure 5.2 is to develop approaches to mitigate any problems your process evaluation has identified. Process evaluations will logically reveal what is working well and what is not in implementing a financial capability program. Because process evaluations tends to be formative rather summative evaluations—that is, occurring as the program is being rolled out—it is important to remember that simply identifying problems or issues is not the ultimate goal; identifying them early on allows implementers to adjust the program in ways that can rectify those problems.

Identifying problems early on allows implementers to adjust ongoing programs to rectify those problems.

In our school-based financial literacy program example, if issues emerge about the material taught or the training of teachers, implementers may need to tweak the material or training accordingly. They may also want to involve at least some of the program's stakeholders in this process, such as some teachers or school directors in our example above. They may be able to help develop suitable and practical solutions to some of your program's challenges.

KEY POINTS

Planning a process evaluation involves the same initial steps as planning for monitoring and for an impact evaluation, beginning with the common process of conceptualizing program components and processes using the results framework model and identifying indicators. After this stage, the difference lies in the objectives of the

evaluation, the measures and methods employed, the type of data obtained, and the outcomes of the analysis. However, further discussions of logistics, analytical methods, and presentation in this Toolkit apply equally to process and impact evaluations.

The core task of designing a process evaluation consists of the selection of the methods for data collection. As mentioned above, data collection can be accomplished by drawing on the program's monitoring system; quantitative research activities; and/or qualitative research activities. The analysis of these data should enable you to develop recommendations for improvements to those areas of the program that experience challenges or demonstrate weaknesses.

Evaluators need to balance the need for data with the burden on participants in collecting such data.

Process evaluations can stand alone, but it is important to recognize that, by design, a process evaluation can provide only limited evidence about program effects, something that is typically the purview of impact evaluations, which are discussed in chapter 6. The strength of a process evaluation is in its ability to provide implementers, practitioners, and other stakeholders with a comprehensive picture of how well the program is progressing relative to expectations. As such, the findings from a process evaluation can be extremely valuable as part of a comprehensive evaluation strategy by providing important contextual information, supporting the interpretation of the results of an accompanying impact evaluation, and identifying specific areas within the structure and management of the program for improvement. Chapter 7 discusses bringing these elements together in greater detail.

FURTHER READING

General

Bliss, M., and J. Emshoff. 2002. "Workbook for Designing a Process Evaluation," Atlanta, GA: Georgia Department of Human Resources, Division of Public Health.

Linnan, L., and A. Steckler. 2002 "Process Evaluation in Public Health Research and Interventions," Jossey Bass Publishers.

Technical

McGraw S. A., Sellers D. E., Johnson C. C., Stone E. J., Backman K. J., Bebchuk J., et al. 1996. "Using Process Data to Explain Outcomes: An Illustration from the Child and Adolescent Trial for Cardiovascular Health (CATCH)," *Evaluation Review* 20: 291–312.

Oakley A., Strange, V., Bonell, C., Allen, E., and Stephenson, J. 2006. "Process Evaluation in Randomised Controlled Trials of Complex Interventions," *British Medical Journal* 332: 413–16.

Saunders, R. P., Evans, M. H., Joshi, P. 2005. "Developing a Process-Evaluation Plan for Assessing Health Promotion Program Implementation: A How-to Guide," *Health Promotion Practice* 6 (2): 134–47.

CHAPTER 6

Impact evaluation

Whereas a process evaluation answers questions about a financial capability program's development and implementation, an impact evaluation is a powerful tool for learning about whether or not a financial capability program is accomplishing the goals and objectives set out for the program. Results from impact evaluations are key inputs in helping funders, policy makers, and other stakeholders determine how successful a program is, and they are also key inputs into conducting cost analyses, which we discuss later in chapter 11.

In this chapter, we first define key terms and concepts related to impact evaluations and then discuss some commonly used evaluation methods that are **not** appropriate for impact evaluations. Then, we introduce the full menu of quantitative methods that are appropriate and available for conducting impact evaluations of financial capability programs. The methods described here are appropriate for both financial education programs intended to increase financial knowledge and for financial behavioral interventions designed to influence financial decision making. We highlight the evaluation methods using examples from the Russia Financial Literacy and Education Trust Fund (RTF) pilot programs and from those published in the economics and development literature. We also discuss practical and logistical challenges in impact evaluation.

Much of the material presented here can be found in other well-written toolkits that present quantitative methods for a general intervention setting; here, we focus on methods and examples specific to financial capability interventions. This chapter is one of the lengthier and more detailed chapters in the Toolkit. In thinking about how to best access the material in this chapter, we recommend reading sections 6.1 and 6.2 for readers who want an overview of impact evaluation concepts. Figure 6.1 at the end of section 6.1 describes the type of evaluation methodology most relevant by program characteristics, and can provide useful guidance on which subsections of this chapter (and other chapters) to explore in more detail. The mathematical background for each methodology described in this chapter can be found in appendix B. Finally, a list of references for further reading is provided at the end of the chapter.

6.1 KEY CONCEPTS FOR IMPACT EVALUATION

An **impact evaluation** is a specific type of evaluation designed to answer a specific question: **What is the causal effect of an intervention on an outcome of interest?** Put differently, what changes in outcomes are **directly attributable** to the intervention alone? The key concern here is being able to tease out whether the financial capability intervention directly causes the changes in the outcomes observed or whether those changes are the result of other factors in the environment. For example:

An impact evaluation separates or helps tease out the effects of the intervention from the effects of other factors.

- Does a school-based financial literacy curriculum cause more savings by students, or do savings rates change for some other reason?
- Does a soap opera's financial capability storyline cause an increase in calls to the debt hotline, or is the increase in calls the result of other factors?
- Does the simplification of loan disclosure information cause low-income consumers to choose less costly loan products, or are changes in product usage caused by other factors?

While causality is a straightforward concept to understand, it can be difficult for a financial capability program to establish, because other factors besides the intervention can affect the financial behavior of program participants.

For example, suppose we observe higher savings rates among participants in a school-based financial capability program. Suppose further that the adoption of the financial capability curriculum is optional and that students in schools who decide to adopt the curriculum come from higher-income families than do students in schools who do not receive the intervention. We might incorrectly conclude that the new curriculum leads to increased savings rates, when, in fact, it was the higher income levels in the participant populations that led to higher savings rates. In this case, different household income levels could affect the results we see.

Things that distort the association between the financial capability program and financial behaviors (in this case, household income) are called **confounding factors**. The aim of quantitative impact evaluation methods is to remove the effect of confounding factors as much as possible to establish a causal link between the financial capability intervention and the financial behavior being studied.

The "counterfactual" is an unobserved state where the program participant did not receive the intervention.

In an ideal imaginary situation, we would observe an individual in two alternate worlds, one in which she received the financial capability intervention and one in which she did not. The alternate world where the intervention was not received is known as the **counterfactual**. Taking the difference between the outcomes in these two imaginary settings would give us the impact of the intervention, because the

only difference between the two settings is the individual's participation in the intervention in one of them and not in the other.

Continuing with the school curriculum example, we would want to compare the world in which a student follows the new curriculum and a parallel world where this same student follows the old curriculum. We would then measure the student's savings rate in both worlds, and the difference in savings would give the causal effect of the financial literacy curriculum on savings.

TREATMENT VERSUS CONTROL

Treatment: those who receive intervention

Control: those who do not

Obviously, we don't have alternative worlds that allow for a clean counterfactual. So the goal of an impact analysis is to rigorously "estimate" the counterfactual using statistical tools. Conceptually, evaluators do this by comparing two groups of individuals: the **treatment group** (defined as participants who receive the financial capability intervention) and the **comparison** or **control group** (defined as similar participants who do not receive the intervention).

An ideal comparison group should:

- Be identical to the treatment group in terms of characteristics, both **observable characteristics** (such as race, gender, education, etc.) and **unobservable characteristics** (such as motivation, preferences, family support, etc.), so that differences in outcomes are not confounded by other characteristics.
- Be expected to react to the financial capability program in the same way as the treatment group. That means that if the intervention were switched and given to the comparison group instead of the treatment group, the impact would be identical.
- Be equally exposed to other interventions as the treatment group. The only difference between the treatment and control groups should be the intervention itself.

TWO KEY QUANTITATIVE APPROACHES

Experimental: randomly assign participants to treatment and control groups

Quasi-experimental: use statistical methods to mimic random assignment

To strengthen the validity of the comparison, we rely on specific statistical methods: (1) **experimental methods**, in which program participants are randomly assigned to the treatment and comparison groups; and (2) **quasi-experimental methods**, in which statistical methods are used to mimic random assignment. The former methods estimate the counterfactual by taking advantage of random measurement error, while the latter methods form control groups with the knowledge that they are not random. Box 6.1 discusses an issue known as **selection bias** that may arise if the treatment and comparison groups are not formed using experimental methods.

In the next few sections, we review the details of a variety of experimental and quasi-experimental methods. Experimental methods include randomized control trials and encouragement design. While experimental methods are the most conceptually straightforward way of forming the treatment and comparison groups, they

BOX 6.1 SELECTION BIAS

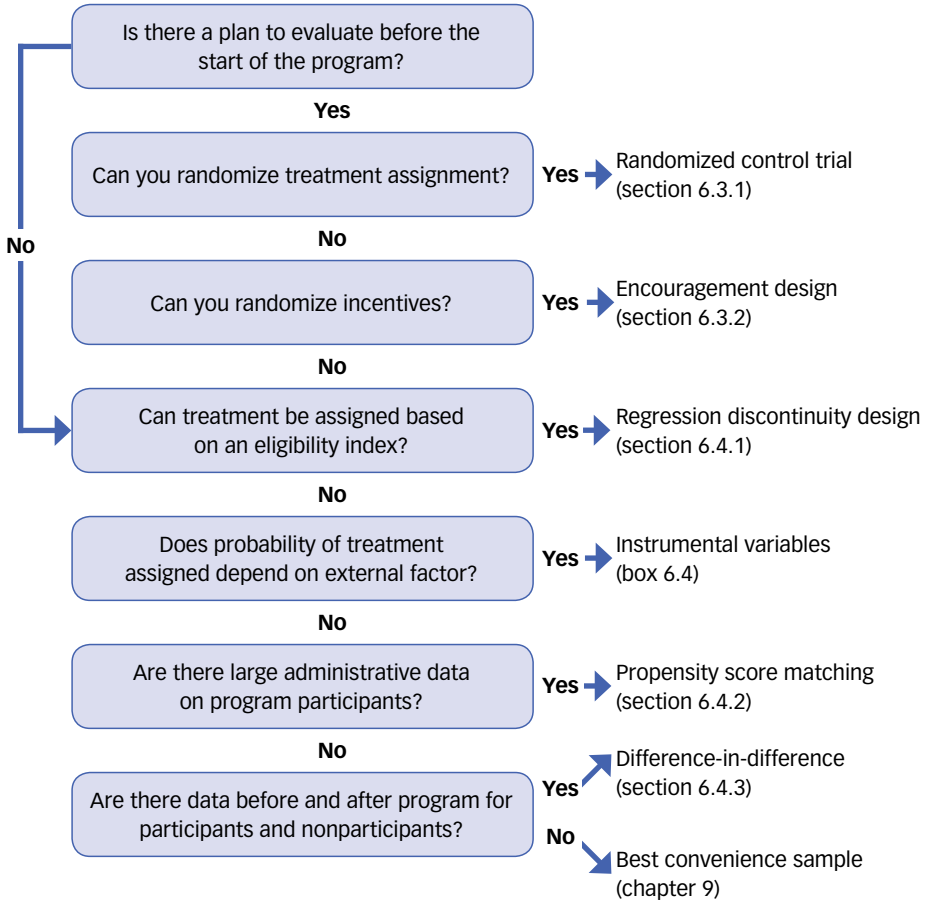
Selection bias occurs when some characteristic(s) of the population causes them to self-select into a group (such as the treatment group), making it difficult to determine whether differences in outcomes are the result of this underlying characteristic(s) or of the program itself. For example, you might see significantly higher financial literacy test scores among people who participate in a financial capability intervention compared to those who did not participate, and credit the intervention for improving test scores. However, because of self-selection, there could be a number of differences between the people who chose to participate in the intervention and those who chose not to. For instance, those who chose to participate might have higher education levels and more concern about their finances than those who did not participate, and these differences may lead participants' test scores to be higher than nonparticipant scores, even without any intervention. Distortions such as this are called selection bias. If selection bias is a factor and not controlled for as best as possible through statistical methods, then it is not meaningful to simply compare the two sets of scores.

are not always feasible; in settings where experimental methods are not feasible, it is important to recognize this fact and employ the most appropriate quasi-experimental method. The quasi-experimental methods described below include regression discontinuity design, propensity score matching, and difference-in-difference.

First and foremost, in choosing an evaluation design, it is wise to consider "the overarching principle…that the rules of program operation provide a guide to which method is best suited to which program and [that] those rules can and should drive the evaluation method, not vice versa" (Gertler et al. 2011). Figure 6.1 presents a decision tree for choosing an appropriate impact evaluation design, starting from the top with Randomized Control Trials and progressing through the list. If one cannot answer "yes" to the question in the figure, then one would move on to the next approach. Importantly, as shown in the figure if there is a not a plan in place to conduct an evaluation prior to the start of the financial capability program, then one can't use either of the two experimental designs. The figure also shows where in this chapter the methodology is described, and the introduction to each methodology in those subsections describes all relevant key terms. If the answer is "no" to all the questions, the final approach is to use the best convenience sample available. Convenience samples are described in chapter 9, as noted in the figure.

Before we discuss the types of experimental and quasi-experimental approaches, we briefly discuss some methods that will not work.

FIGURE 6.1 DECISION TREE FOR CHOOSING IMPACT EVALUATION DESIGN



6.2 A CAUTION AGAINST FALSE METHODS

Determining the best possible estimate of the counterfactual is at the heart of quantitative evaluation design. But not all quantitative evaluation designs are created equal. Some common research designs are relatively easy to put into practice but **can't** provide good estimates of the causal intervention effect, because the counterfactual is not sound:

Some common research designs are easy to put into practice but cannot provide good estimates of causal effect.

- Comparisons of outcomes before and after the program for those who participated
- Comparisons of outcomes between voluntary participants and nonparticipants.

It may be tempting to compare the outcomes for participants before and after the program, in what is called a **pre-post design**. However, in this case, there is no way

BOX 6.2 WHAT IS THE ROLE OF QUALITATIVE RESEARCH IN IMPACT EVALUATION?

In the evaluation literature, we often use the technical term "impact" specifically to refer to a quantifiable change in an outcome measure that is attributable to an intervention, and "impact evaluation" to refer the statistical estimation and testing of causal effects. It is important to note that impact evaluation as defined cannot be purely qualitative. Even if we compare focus group discussions about outcomes collected from two randomly assigned treatment and comparison groups, qualitative methods cannot conclusively establish causal effects, in the sense of yielding estimates that can be subjected to a rigorous test. Many evaluators therefore typically think of impact evaluation as an entirely quantitative exercise. However, it is important to understand both the strengths and limitations of qualitative analysis in this context. Qualitative assessment of outcomes provides important context and explanation for why something happened. As such, it can and should be a part of any impact evaluation as an important complement (but not a substitute for) to quantitative impact evaluation methods. Furthermore, in any summative evaluation, it is important to combine quantitative and qualitative methods and to integrate elements from process evaluations and monitoring into impact evaluations as appropriate.

to tell whether changes in outcomes result from the intervention or any other event that occurred at the same time. This is particularly true if the outcomes of interest are time-sensitive.

For example, schoolchildren's financial capability scores may increase the year after a financial capability curriculum is administered. However, even without the curriculum, we might see natural increases over time because of other factors, such as the fact the children are growing older and, as a result, are developing mathematical ability and getting a more general exposure to financial matters. There is no way to separate these effects without a comparison group. To get an estimate of the program impact in this setting, you would have to assume that every factor that affects the financial behaviors stayed constant during the time the intervention was in place. While evaluators can control for many time-varying factors that could affect outcomes, in practice it is impossible to collect enough data to capture all relevant measures exhaustively.

The problem with comparing voluntary participants to nonparticipants is selection bias, which we discussed above in box 6.1. It is not enough to control for differences that are observable, because measures of characteristics such as ability and motivation that determine participation are usually not available to the evaluator. If the individuals who chose to participate in the financial capability intervention are different from nonparticipants on any such characteristic that is not captured, then evaluators cannot clearly attribute the change in outcomes to the intervention.

While these two methods are not useful for estimating the causal impact of a program, information collected from program participants can and should be used for monitoring and process evaluation, as long as the limitations of this information are understood.

6.3 EXPERIMENTAL EVALUATION DESIGNS

As noted above, experimental evaluation designs are those in which program participants are randomly assigned to the treatment and comparison groups. We look at two types here: Randomized control trials (RCTs), and encouragement designs.

6.3.1 Randomized control trial

When conducted properly, an RCT is the best possible methodology for ensuring a valid counterfactual—it is considered the "gold standard" in evaluation. In this method, evaluators form the treatment and comparison (or control) groups by randomly assigning the financial capability intervention to a fraction of the eligible participants. Random assignment means that eligible participants are equally likely to be placed in the treatment group or the control group. As long as the group of participants is large enough, the groups will be statistically identical.

RCTs are considered the "gold standard" in evaluation.

One positive consequence of randomized control design is that every eligible participant has an equal chance of receiving the financial capability intervention. This means that the design provides intervention administrators with a fair and transparent rule for allocating scarce resources equally among program participants.

PROGRAM CONDITIONS FOR SELECTING THIS EVALUATION APPROACH

When does it make sense to conduct an RCT? Evaluators can use this approach if:

- The financial capability program's conditions make it **feasible** to assign participants to a treatment and control group.
- Evaluators can randomize **prior to the start of the program**.
- The program has a **large enough number of participants** in both the treatment and comparison group to allow for meaningful statistical analysis.
- It is easy for participants to **comply with the assignment**.
- Evaluators have the **institutional support** needed to maintain and monitor the experiment.

For example, in evaluating a school curriculum related to financial capability, it is feasible to randomly assign a large group of school children prior to the start of the school year to classrooms that receive the new curriculum (treatment group) and old curriculum (control group). Because the students receive the intervention in a school setting, school leaders can ensure that the students receive the correct curriculum, and financial capability tests can be administered to both the treatment and control group students before and after the curriculum is taught.

In contrast, if policy makers are interested in evaluating a soap opera's financial education storyline, they can't have evaluators conduct an RCT of the storyline, because there is no method that can reasonably enforce the assignment of the treatment and control groups.

Randomized assignment can be used when there is an oversubscription to an intervention or when resources are limited in delivering the intervention to the entire population. Randomization can also be used to phase in an intervention over time to a large population by introducing the intervention one small group at a time. In this case, the participants who have not yet received the intervention will form the control group.

INTERNAL VERSUS EXTERNAL VALIDITY

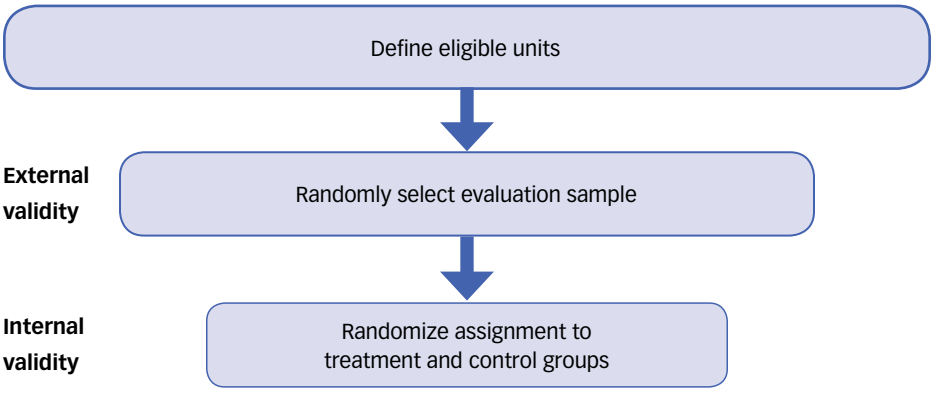
Internal: all confounding factors have been removed so comparison group is true counterfactual

External: evaluation results can be generalized to population of all eligible units

IMPLEMENTATION

Randomization should be conducted in three steps to ensure that the confounding factors have been removed (internal validity) and that the results can be generalized to entire population of eligible individuals (external validity). To elaborate, **internal validity** means that all confounding factors have been removed and that the comparison group represents the true counterfactual; therefore, intervention effects can be considered causal. **External validity** means that the impact evaluation results can be generalized to the population of all eligible units. The three steps are shown in figure 6.2 and discussed below.

FIGURE 6.2 THREE STEPS IN RANDOMIZED ASSIGNMENT



- **Step 1: Define eligible units.** The eligible units are defined as the population you are interested in studying in the financial capability program being implemented, such as school children, television viewers, or low-income citizens.
- **Step 2: Select evaluation sample.** If the eligible population is small, then it can serve as the evaluation sample. However, if the eligible population is too large, then evaluators need to randomly select a representative sample of

units for the evaluation. The size of the evaluation sample is determined by a statistical calculation known as a **power calculation**, which is discussed in chapter 9 and in appendix B.

- **Step 3: Randomize assignment to treatment and control groups.** The final step is to randomly assign units to the treatment and control groups. Typically, the randomization is conducted as a lottery draw and performed by computer software or by flipping a coin. It is important to establish a rule for how people will be assigned to the treatment **before** the random numbers are generated (or the coins are flipped). This ensures that the randomization is not invalidated by researcher interference.

A key concern for RCTs is the decision about what level to perform the randomization at. Random assignment can be done at the individual, household, school, community, or regional level. This decision depends partly on how the program will be implemented.

For example, for school-based programs, randomly assigning schools to a treatment or control group is logical, whereas for the purposes of financial literacy mentoring, an individual or community-based randomization may be more fitting.

The randomization level also depends on whether the evaluators are interested in different groups of eligible units. For example, men and women respond differently to a financial mentoring program. In this case, the eligible population can first be divided into subgroups (by demographic characteristic such as gender, age, income,

BOX 6.3 RTF IN PRACTICE: DESIGNING RANDOMIZATION TO AVOID SPILLOVERS

In Uganda, researchers are performing a randomized evaluation of KASSIDA, a business education program provided to small workshops in Uganda. Owing to the physical clustering of workshops, the randomization is at cluster rather than individual level. Researchers hope to mostly eliminate the direct copying of new techniques learned in training—the first source of potential spillover effects—because workshops operating side-by-side are either all receiving training or part of our control group. However, interestingly, they also hope to directly assess the size of the spillover effects by mapping the full spectrum of inter-relations between workshops that are part of our study (both within and across treatment and control groups) and also with those outside our sample. To assess the spillover effects, outcomes of interest will be tracked for up to five network contacts for both treatment and control groups, and researchers will compare changes in performance for the contacts of each group that are outside the study sample.

BOX 6.4 RTF IN PRACTICE: EXAMPLE OF AN RCT IN BRAZIL

One of the RTF pilot programs implemented a school-based financial education curriculum in Brazil. The eligible population for this study was identified as students in their last two years of secondary school. In the pilot program, 891 schools in 5 states and the Federal District were identified as the evaluation sample. Because these states and schools were not selected at random from all states and schools in the country, the external validity of the evaluation is not guaranteed.

The random assignment was stratified by state and by how much financial inclusion there was in the municipality where the school is located. Within each stratum, schools were matched into pairs by student and neighborhood characteristics, such as number of students and teachers in the high schools, dropout rates, graduation rates, municipality GDP per capita, and municipality savings per capita. Within these pairs, one school was randomly assigned to treatment and the other school to the control group. The table shows the total number of schools and students in each group.

Results of school-level randomization

	CONTROL GROUP		TREATMENT GROUP	
STATE	SCHOOLS	STUDENTS	SCHOOLS	STUDENTS
Ceará	60	1,959	61	2,012
Distrito Federal	29	1,112	32	1,105
Minas Gerais	15	371	14	360
Rio de Janeiro	136	3,752	134	3,513
São Paulo	174	6,037	168	5,471
Tocantins	17	514	17	505
Total	431	13,745	426	13,236

The table below compares characteristics between the treatment and control groups.

Comparison of characteristics

VARIABLE	TREATMENT	CONTROL	DIFFERENCE
Female	0.562	0.549	0.013
Age ≤ 13	0.003	0.004	–0.001
Age = 14	0.013	0.009	0.004
Age = 15	0.150	0.148	0.002
Age = 16	0.486	0.510	–0.024*
Age = 17	0.227	0.221	0.006
Age ≥ 18	0.121	0.109	0.012
Baseline financial test results			
Number of points	15.660	15.549	0.111
Percentage right answers	0.476	0.472	0.004
Level of proficiency	50.149	49.799	0.350
Autonomy score	50.028	50.081	–0.053
Intention to Save score	49.919	50.011	–0.092

Note: * indicates significance at 5% level.

While only 12 characteristics are displayed in the table above, the researchers actually checked 73 characteristics. Of the 73 characteristics tested for equality, seven were statistically significantly different between the treatment and control groups. Since the seven characteristics comprise 9.4 percent of the tested characteristics, the randomization was successful, and the comparison group is a good estimate of the counterfactual.

etc.), and randomization can be performed within the subgroups. This method is known as **stratification**. Stratification is useful for cases where the demographic characteristics are not evenly distributed in the population but where policy makers are interested in the effects of the intervention on subgroups of the population.

Stratification is useful where demographic characteristics are not evenly distributed in the population but where policy makers are interested in the effects on subgroups.

Another important consideration for deciding the level of randomization is the trade-off between having a large enough sample to estimate true differences in the treatment and control group while avoiding "contamination" of the financial capability intervention. A higher level of randomization (e.g., at the school or region level) can threaten the ability to draw conclusive results from the evaluation if there are not enough observations that form the treatment and control groups (i.e., too few schools or regions).

However, randomization at the individual level is more likely to be threatened by contamination or **spillovers** (when participants in the control group are affected by the treatment). In the financial capability setting, spillovers are typically in the form of what are called learning and imitation effects. Learning effects may be possible if students assigned to receive a financial capability program talk to their friends in the control group about what they learn. Alternatively, imitation effects may be an issue, especially because financial capability programs are highly concerned with changing behavior. In such a case, students in the control group may not learn anything themselves, but they may simply emulate their friends' behavior. In either case, practically speaking, the experiment is contaminated.

Randomization at the individual level can be threatened by spillovers, where the control group is affected by treatment.

If spillovers are expected to be local, such as students in the treatment and control groups talking to one another, then evaluators can randomize at the group level and estimate the total program effect on the group. In this example, randomizing schools to the treatment and control groups will avoid the spillover issue, as long as students from different schools do not discuss the intervention. Therefore, where spillovers are likely to be important, experiments should be specifically designed to account for them.

After randomizing, it is important to check whether the randomization resulted in two groups that are, in fact, similar. This is typically accomplished by testing whether the average characteristics of the control and treatment groups are statistically the same between the two groups.

All the data collected on the treatment and control groups should be included in this check. In addition to characteristics such as age, gender, education, etc., all the outcome variables of interest, collected before the financial literacy intervention, should also be included in the check. For statistical reasons, we would expect that even when the treatment and control groups are statistically identical, a small percentage of the characteristics will be different between the two groups. As long

as differences occur in only 5–10 percent of the tested characteristics, we can consider the randomization process to be successful, and the comparison group will be a valid estimate of the counterfactual.

LIMITATIONS

While randomized control design is sometimes called the gold standard for quantitative impact evaluation, an RCT can be costly to implement and maintain over an entire evaluation, and implementation is rarely perfect. While it is simple to evaluate the program if implementation is perfect, analysis becomes more complex if there are issues with contamination, if participants drop out of the treatment or control groups (attrition), or if there are other similar concerns that jeopardize the random assignment. (See, for example, the next section.)

There are many contexts in which it is simply not logistically or politically possible to randomize treatment and control group assignment. Box 6.5 discusses a version of RCT that is possible when it is not politically feasible to withhold the program from everyone who is eligible. It is important to keep in mind that even in cases where randomization is not possible under any circumstance, there are still options for rigorous impact evaluation methods.

BOX 6.5 RANDOMIZED PHASE-IN

In some situations policy makers may be interested in the impact of a financial capability program, but it may be impossible to withhold the program from everyone who is eligible to receive the program, especially if the program spans multiple years. In these cases it may be possible to roll out the program over time to randomly selected groups of the population. This variation of a randomized control trial is known as **randomized phase-in**.

To implement a randomized phase-in, the eligible population is randomly assigned to more than two groups. Over time, each group is individually assigned to receive the intervention, and will be considered the treatment group for a span of time. The program participants that have not yet received the intervention will be used as the control group for that time period. Over the course of multiple years all of the groups will receive the treatment at some point.

A randomized phase-in design is naturally well suited for long-term programs that are expected to roll out over multiple years. Because every eligible unit eventually receives the intervention, this method is politically advantageous. However, concerns of contamination and attrition that were discussed in traditional randomized control trials above are magnified in a phase-in design setting because the program spans multiple years. Also, the long-term impact of a financial capability intervention cannot be estimated because eventually every eligible unit will receive the treatment. Finally, a phase-in design requires a dedicated evaluation team who is willing to monitor the roll-out of the program for the entire life of the project, and therefor this method could be more costly.

6.3.2 Encouragement design

In many financial capability programs, it is not practically possible to exclude anyone from taking part in the intervention. In low- and middle-income countries (LMICs), classic examples are social marketing or mass media interventions. For example, it is very difficult to ensure that large numbers of people are kept from watching a broadcast television show. But it may still be possible to use randomization.

In such cases, evaluators can use a method called encouragement design (or randomized promotion) and randomly assign who is encouraged to participate in the intervention, instead of randomly deciding who receives the intervention. The promotion could be an information campaign on the benefits of a savings or it could be incentives, such as small gifts or prizes for signing up to participate in a financial capability workshop.

To estimate the causal impact of the program, evaluators compare the outcomes of the promoted group with the nonpromoted group, adjusting for the difference in participation across the groups. This method is one application of a larger class of techniques known as **instrumental variables (or IV) estimation**. IV estimation is a technique commonly used to identify causal effects in nonexperimental settings (see box 6.6 for a detailed explanation of instrumental variables).

Encouragement design is best suited for programs with voluntary enrollment or programs with universal coverage, where enrollment into the program cannot be randomized because of those features. Unlike a traditional RCT, we don't expect that all program participants encouraged to participate in the financial capability intervention will actually do so. Also, some people assigned to the nonpromoted group may decide to participate in the program anyway. The key is that the promotion will encourage some randomly selected eligible people who would not have considered participating in the financial capability intervention to participate; thus, the rate of participation in the promoted group will be higher than that in the nonpromoted group.

It is important to keep in mind that if an encouragement design is used, the causal impact of the program can't be evaluated by comparing the financial capability outcomes of those who participated in the program to those who did not participate. The people who choose to participate in the program may be different from the people who choose not to participate, and these differences can drive the differences in outcomes. Instead, to estimate the causal impact of the program, you should compare the outcomes of the promoted group with the nonpromoted group. Since these two groups were formed randomly, they are statistically identical, and the difference in their outcomes is the best estimate of the causal impact of the program. For technical details on the different types of estimates that can be measured, see box 6.5.

BOX 6.6 INSTRUMENTAL VARIABLES

Instrumental variables estimation can "correct" the estimates of the intervention for having nonrandom treatment and control groups under certain conditions. While there are many different applications where the instrumental variables methodology is appropriate, they share one common feature—a feature that is present in encouragement design—they are examples of situations where randomization affects the **probability** that the participants receive the treatment, not the actual treatment itself.

In general terms, in attempting to estimate the causal effect of a program on an outcome in the absence of randomization, an **instrument** is a factor that affects the outcome only through its effect on the program. Underlying this setup are two characteristics that are required for a good instrumental variable:

- The instrumental variable must be correlated with program participation.
- The instrumental variable may not be correlated with outcomes (except through program participation) or with unobserved variables.

In analyzing the impact of a program under encouragement design, the **assignment of the encouragement** is used as an **instrumental variable** for actual enrollment. Specifically, the instrumental variable is an indicator variable for whether the participant received the encouragement to participate in the intervention. This is a good instrumental variable, because those who receive the encouragement are more likely to participate in the intervention (correlation with program participation), and receiving the encouragement only affects the financial capability outcome through the act of participating in the intervention (uncorrelated with outcome).

Instrumental variables are an appropriate tool for both prospective and retrospective evaluations. It can be used in any setting where the probability of being assigned to the treatment group depends on an external factor and is independent of the outcome of interest. Another example where the technique is used is a natural experiment where "the forces of nature or government policy have conspired to produce an environment somewhat akin to a randomized experiment" (Angrist and Krueger 2001). Natural experiments can be existing policy differences, or changes that affect some jurisdictions (or groups) but not others, or unexpected changes in policy in the local area.

PROGRAM CONDITIONS FOR SELECTING THIS EVALUATION APPROACH

Encouragement design is best suited for programs with voluntary enrollment or universal coverage where program enrollment cannot be randomized because of those features.

An evaluation using encouragement design is possible if:

- **It is not possible or desirable to exclude anyone from participating in the program**; alternatively it is appropriate when it is feasible for all eligible units to participate in the program at the same time, but budgetary reasons prohibit an evaluation of all eligible units using a traditional RCT.
- **The program is not already popular**; otherwise, it will be difficult to find differences in the promoted and nonpromoted groups.

- **The promotion device is very effective**; whatever is used should have effects that are large enough to significantly affect the likelihood that people will participate in the financial capability program.
- **The promotion does not affect the outcome of interest directly.** For example, if the outcome of interest is savings and the promotion is a cash prize, then there may be higher savings by the promoted group simply because they received the promotion. In this case, an alternative form of promotion could be a gift card that cannot be converted to cash.
- **A large number of people are expected to participate.** In an encouragement design, the promoted group needs to have a substantially higher enrollment rate than the nonpromoted group to permit causal inferences to be drawn; thus, the number of people participating in an encouragement design study should be larger than an RCT.

IMPLEMENTATION

The procedure for carrying out an encouragement design study is just like the procedure described for an RCT in section 6.3.1, except that it is the promotion rather than the treatment itself that is randomized. First, the eligible population needs to be defined. Next the evaluation sample should be selected, keeping in mind that this sample may need to be substantially large. Finally, the promotion should be randomly assigned to a fraction of the evaluation sample. The randomization level should be considered to account for program conditions and spillover, and if subgroups of the eligibly population are of interest, stratification should be used. As with the tradi-

BOX 6.7 RTF IN PRACTICE: EXAMPLE OF ENCOURAGEMENT DESIGN, SOUTH AFRICA

The South Africa entertainment education program is an example of encouragement design in the context of financial capability. This program embeds a financial capability storyline in the South African soap opera *Scandal!* The evaluation examines the impact of this intervention on viewers' knowledge about debt management and their behavior, specifically whether they seek assistance from a local debt hotline. The storyline focuses on sound financial management and how to get out of debt.

Because viewing the soap opera cannot be randomized, evaluators used an encouragement design. The encouragement came in the form of financial incentives. In a clever twist on the straightforward encouragement design, both the treatment and control groups received an encouragement, but the treatment group was encouraged to watch *Scandal!*, while the control group was encouraged to watch a different soap opera that airs at the same time as *Scandal!*. To receive the incentives and to check the implementation of the methodology, the intervention participants were required to answer questions on the content of the soap opera they were encouraged to watch.

tional RCT, the implementation should be checked to ensure that promoted and nonpromoted groups are statistically identical.

LIMITATIONS

Encouragement design causal estimates are not externally valid.

Again, like RCTs, encouragement designs have limitations. First, the encouragement must be effective in increasing program participation—an ineffective encouragement will prevent a comparison of the treatment and control groups. This can typically be overcome by pilot-testing the encouragement to ensure that it is effective. Second, it is important to remember that the causal estimate of the program from an encouragement design method is a local average treatment effect (as described in box 6.8), and, as such, the program effect estimates are not externally valid.

BOX 6.8 A NOTE ON TREATMENT EFFECTS

The term **treatment effect** is another way of saying causal program effect. The term originates from the medical literature where treatment referred to a new drug or surgical procedure, but it is used more generally today to refer to the effect of any type of intervention. Depending on the ability of the evaluation design to control for confounding factors, there are a number of types of treatment effects that can be identified by the evaluation. An RCT identifies what is called the **Average Treatment Effect (ATE)**, which measures the difference in average outcomes between units assigned to the treatment and units assigned to the control group.

When evaluators use an encouragement design, they cannot measure the ATE. One effect of interest in this case is the comparison between the outcomes of the group originally assigned to treatment with the group originally assigned to comparison. This comparison will yield the **"intention-to-treat" estimate (ITT)**, because it compares the outcomes of those who were intended to be treated with those who were intended to not be treated. In some cases, this measure can be important, especially if policy makers cannot enforce participation in the program.

However, policy makers are typically interested in the impact of the program on the participants who actually receive the treatment, and the ITT estimate is not the correct measure for this purpose. This is when the instrumental variables method is useful, because it will correct the error introduced by the nonrandom composition of the treatment and control groups in the ITT estimate. It allows evaluators to estimate the impact of the intervention on participants to whom the treatment was offered and who actually enrolled. These estimates are known as the **Local Average Treatment Effects (LATE)** and measure the average effect of the intervention on participants who were induced to take part in the intervention because they were assigned to the treatment group.

6.4 QUASI-EXPERIMENTAL EVALUATION DESIGNS

There are many options for evaluating financial capability programs when randomized assignment is not possible. As mentioned above, one can use "quasi-experimental" methods; in this case, while the treatment and control groups are not formed randomly, statistical methods are used to mimic random assignment. Some of these methods form estimates of the counterfactual by taking advantage of random measurement error, while others form control groups with the knowledge that they are not random. Research has shown that quasi-experimental techniques can provide reliable results when the selection process is known and measured and/or when the treatment and comparison groups are matched on at least pre-intervention measures of outcome.

In this section, we look at a number of quasi-experimental designs: regression discontinuity design (RDD), propensity score matching (PSM), and difference-in-difference (DID).

6.4.1 Regression discontinuity design

In many instances, evaluators may not be able to plan for an evaluation at the start of the financial capability program, thus ruling out the use of an encouragement design. However, it may still be possible to form a valid comparison group when a program relies on certain rules to distinguish between potential participants and nonparticipants. For example, there are many financial capability interventions that rely on ranking participants based on an index, such as age, credit score, or test score, and then assigning the program based on a cutoff in the rankings. By making use of these rankings and cutoffs, it is possible to avoid selection bias in the evaluation.

BOX 6.9 EXAMPLE OF RDD IN PRACTICE

One example of RDD is a published study by Jacob and Lefgren (2004) that examines the impact of remedial schooling (summer school) on student outcomes. In 1996, Chicago Public Schools instituted a policy that placed students in summer school based on their performance on a standardized test. Students who scored below a threshold value (where the threshold depended on their grade level and other factors) were placed in summer school, whereas those who scored above the value were not placed in summer school. This policy resulted in a highly nonlinear relationship between the test score and the probability of attending summer school. By comparing the outcomes of students just above and just below the threshold, the authors get an estimate of the impact of the intervention.

The idea behind RDD is that participants who score just above and just below the threshold are very similar to one another. Therefore, if the financial capability program is assigned to those who score below the threshold, those who are **just below** the threshold will be the treatment group, and those who score **just above** the threshold will be in the comparison group. For example, if participation in a financial literacy training program is determined by scoring in the bottom 25 percent of a screening exam, the exam scores could be used as the eligibility index, and those who score just below the 25th percentile will be in the treatment group, while those who score just above the 25th percentile will be in the comparison group.

The score that is used to assign participants to the treatment and control groups is known as the **eligibility index**. This measure may be used directly to decide assignment, or it can be used in combination with other measures to form the **control function**, which takes multiple measures and combines them into a single index. The RDD methodology is based on the statistical fact that there is error in the measurement of the eligibility index or control function and that the small error that places participants just above and just below the threshold implies that assignment to the treatment group can be considered close to random. Because the participants around the threshold are similar, the comparison group is a strong estimate of the counterfactual. Thus, where RDD is possible, it provides the best alternative to randomized control design for examining the causal effect of financial capability interventions.

The RDD method takes advantage of existing rules in the program and allows for evaluating the program without changing program design. Because it does not rely on random assignment, the RDD method can be used as a retrospective evaluation tool. Thus, it may be politically more appealing.

The RDD method takes advantage of existing program rules and allows for evaluating the program without changing program design.

There is one basic extension of RDD that accounts for a case where the assignment variable is not perfect at placing participants in the treatment group at the threshold. This is known as "fuzzy" RDD. In this case, instead of determining treatment status, the control function determines the **probability** of treatment. The solution to this scenario is to use the threshold-based assignment as an instrumental variable to recover the treatment effect (see box 6.6 for a detailed explanation of instrumental variables).

PROGRAM CONDITIONS FOR SELECTING THIS EVALUATION APPROACH

An RDD method can be implemented if:

- **Program assignment is based on an index** that is continuous, such as age or a test score on which it is possible to rank the participants in the program.
- **There is a cutoff score that defines eligibility.**

- **There is little or no scope for manipulation of the program rules.** Major stakeholders (governments, policy makers, participants, and researchers) should not have control over exactly where program participants place in terms of the eligibility index cutoff. The relationship between the assignment variable, the formation of the control function, and the threshold should be nontransparent to prevent the manipulation of the treatment and control groups by program participants.

In the case of financial capability programs, this design can be used in very natural settings. For instance, if individuals are admitted to financial capability programs based on performance on a knowledge test, comparing individuals just above and below the cutoff point will give an estimate of the program impact. Another example may be to compare individuals assigned to counseling or other interventions based on their credit score falling below a certain limit to a comparison group of individuals who are just above the qualifying mark. If desirable, it is possible to combine this test score with other variables (e.g., age, income, gender, etc.) to form the control function.

IMPLEMENTATION

There are multiple steps in implementing an RDD methodology. First, the eligibility index needs to be selected. The eligibility index can be any continuous score that is used to assign participants to the program. As noted above, it is possible to combine multiple measures into a single control function. Once the assignment variable has been selected, the cutoff point needs to be selected as well.

The next step is to establish the analysis sample, which determines how "close" to the threshold the score of the participant must lie for that participant to be included in the analysis. It is important to strike a balance between the need for a large sample size and concern that the treatment and comparison groups are valid. If the boundary for the analysis sample is too close to the threshold, there may not be enough observations to draw conclusions about the effectiveness of the intervention. However, the more observations that are included away from the threshold, the more likely it is that the comparison group is different from the treatment group and, therefore, is not a valid estimate of the counterfactual.

The RDD method makes sense where there is little or no scope for manipulation of program rules.

Typically, the analysis sample is created by assigning a weight to all the participants, with a lower weight assigned to participants who are further from the threshold. To draw conclusions about program effectiveness using RDD, it is important to have an adequate sample size close to the threshold. The required sample size can be calculated using a power calculation. (For more details, see appendix B.)

As with the methods we have discussed so far, it is important to ensure that the method has been properly implemented. The first step is to graph the eligibility

index of the sample on the x-axis against the outcome measure of interest on the y-axis and then examine whether there is indeed a discontinuity at the threshold of the eligibility index. It is also important to compare the characteristics of the treatment and control groups to establish that the two groups do not differ notably along observable dimensions. After the results of the intervention are available, it is also important to check that using a different analysis sample (by assigning different weights) does not have a significant impact on the results. Typically, this is done by checking that the results are similar using a variety of weighting functions.

LIMITATIONS

The results from an RDD analysis—just like those in an encouragement design—can't be generalized to the all the participants in the program because the method only analyzes the population near the threshold. Thus, the results from RDD are also local average treatment effects (see box 6.8 for a review of treatment effects). In some cases, it is this local population that is of particular interest, especially if one of the reasons for the evaluation is to consider expanding the program at the margin (to a few more participants near the threshold). But the RDD methodology can't answer the question of whether the program should exist at all, because the analysis does not include the participants who are far from the threshold.

RDD results are also not externally valid, because the method only analyzes the population near the threshold.

In addition, because the analysis focuses on program participants close to the threshold, it requires a large sample size around the threshold.

Finally, especially in LMICs, it is important to consider whether the eligibility rules will be enforced. While "fuzzy" regression discontinuity methods can be used in some instances where enforcement is weak, if there is concern that the eligibility rules will be completely ignored or manipulated by those involved in the program or evaluation, then the RDD method can't be used to evaluate the impact of the program.

6.4.2 Propensity score matching

When random assignment, random encouragement, or regression discontinuity methods are not possible, matching methods can be used as an alternative evaluation method. While these methods were popular in the past, they have been eclipsed by the methods described above, which are considered more robust.

Matching methods can suffer from the curse of dimensionality—too many characteristics to match.

Matching methods are an intuitive implementation of the following idea: We can't observe the same individual in two worlds at the same time, but we can find individuals who have very similar characteristics and assign them to the treatment and comparison groups based on their observable characteristics.

For example, we can compare a 20-year-old woman in the treatment group with a 20-year-old woman in the comparison group and estimate the impact of a financial

capability workshop by comparing the outcomes for these individuals. This methodology assumes that it was random chance that only one of these two very similar persons received the financial capability intervention. The matching method mimics randomization by trying to create an observational analogue of a randomized experiment and by discarding observations without a good match.

From a practical perspective, we should match individuals on the determinants of program participation. If the list of relevant determinants is very large, or if each characteristic takes on many values, it may be hard to identify a match for each person in the treatment group. As the number of matching characteristics or dimensions increases, evaluators may run into what is called **the curse of dimensionality**.

For example, if two characteristics (age, gender) explain program participation, it may easy to find matches for all treatment participants, where women are matched to women and then ages can be approximately matched. But if in addition there is a need to match on income, education level, ethnic group, and other characteristics, the chance to match exactly on all those characteristics diminishes very quickly. As such, it may be difficult to match two 20-year-old women from a specific minority group who have low incomes and just primary education.

Fortunately, there is a solution to the curse of dimensionality. An extension to simple matching, known as **propensity score matching (PSM)**, takes a number of measures and combines them into a single score, similar to the control function described in the RDD section. The score is known as the propensity score, and it is a number between 0 and 1 that represents the predicted probability of participating in the intervention. Rather than matching on multiple characteristics, participants are instead matched only on the propensity score. Intuitively, people with a similar propensity score have a similar probability of being in the treatment group. Maybe these individuals are not exactly the same on the observed characteristics, but overall they are equally likely to participate in the program.

PSM deals with the curse of dimensionality by combining a number of measures into a single score.

As with the simple matching described above, we can estimate the impact of the project by comparing the outcomes of individuals with similar propensity scores. By collapsing the many matching characteristics into a single score, the propensity score addresses the dimensionality problem. However, note that PSM maintains the strong assumption that all characteristics excluded from the matching process (because they are unobserved or unmeasured) are either not correlated with the outcomes or do not differ across participants and nonparticipants. Because of this assumption, PSM is not considered a strong research design and should only be used in combination with other methods or in cases when other experimental or quasi-experimental methods are unavailable.

PROGRAM CONDITIONS FOR SELECTING THIS EVALUATION APPROACH

The requirements to conduct a PSM evaluation revolve around the data that were collected. In particular, PSM can be used if:

- **There are data available on a large number of participants and nonparticipants.** Because individuals need to be paired on fundamental characteristics that are unaffected by the intervention, this requires a significant amount of data.
- **The data contain a large number of characteristics** that are either unaffected by the financial capability program (e.g., age or gender) or measured before the program was implemented (such as outcomes at baseline, if available).

While it is not a necessary condition, most PSM evaluations are retrospective, which means that they take place after the program has been implemented. It is possible to implement a PSM evaluation with a single cross-section of data that was collected after program implementation as long as program participants are identified.

IMPLEMENTATION

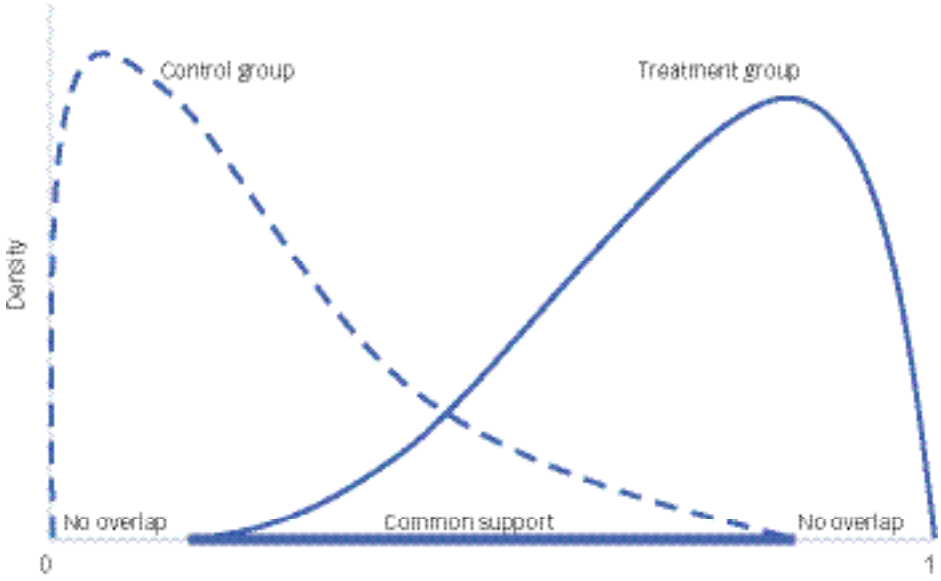
An important feature of PSM is that not all participants and nonparticipants are included in the analysis sample. This is because some members of the treatment group may not have a good match in the comparison group or vice versa.

Because PSM is conducted after the program was implemented, the evaluation implementation pertains to how the data are analyzed. PSM is implemented in six steps:

1. Identify treatment and comparison group participants and combine into one sample.
2. Estimate the probability that each person participates in the program using the individual-level matching characteristics. This estimation is done for the combined sample and uses only baseline characteristics that are not affected by the program. This step yields the propensity score.
3. Restrict the analysis sample to those individuals in the treatment and comparison groups who have overlapping propensity scores.
4. For each participant, find one (or many) nonparticipants who have a similar propensity score.
5. For each of these "matches," calculate the difference in the outcome of the participants and nonparticipants.
6. Obtain the program's average treatment effect (for all participants) by taking the average of the difference in outcomes.

The most straightforward manner for finding the analysis sample is to graph the propensity scores for the treatment and comparison groups. Figure 6.3 displays a simplified example. The participants in the treatment and comparison group with propensity scores in the area labeled "common support" will comprise the analysis sample. The outcomes of the remaining participants, labeled "No overlap," will not be used in the analysis.

FIGURE 6.3 PROPENSITY SCORES OF TREATMENT AND COMPARISON GROUPS



For individuals who have propensity scores in the common support area, we need to identify the best possible match. Intuitively we would like matches to have similar propensity scores. There are many approaches to specifying what "similar" means that will lead to somewhat different matches. There is no universal best approach, and it is good practice to assess the sensitivity of the results by trying several methods. Similarly, we can choose to match each participant with one or many nonparticipants. The number of matches can also be investigated in a sensitivity analysis.

Steps 5 and 6 (above) can be implemented in a regression framework that can also account for other individual characteristics. (The interested reader may also see Jalan and Ravallion (2003) and Gertler et al. (2011).)

BOX 6.10 EXAMPLE OF PSM IN PRACTICE

While none of the RTF pilots use PSM to evaluate program impacts, Agarwal et al. (2010) evaluate a voluntary long-term program to assist low-income households to improve their financial management to support mortgage choices and homeownership in Indianapolis, United States. Participants are poorer and have worse credit histories than the average borrower in this area, reflecting the successful targeting of the program. Importantly, they are also likely to be more motivated than the average low-income household: the program is voluntary and requires a substantial commitment. A comparison of the participants and nonparticipants could be flawed for both reasons. The authors use PSM to identify nonparticipants that are comparable to participants, and then compare the average outcomes for these groups.

Like RCTs, the characteristics of the treatment and comparison groups should be carefully compared to ensure that the matching was successful. Comparisons of multiple characteristics are described in detail in the RCT section 6.3.1.

LIMITATIONS

PSM has its share of limitations. A practical challenge is, as noted above, the need for a large, detailed data set of participants and nonparticipants. The data set must contain sufficient information to adequately estimate propensity scores. Often the data available for matching are collected only once, and in many instances, this information is collected after the program has been implemented. Administrative data often don't have enough detail to implement PSM. Also, as discussed above, not all individuals are good matches and may be discarded in the matching process. A larger data set makes it more likely that the final analysis sample has a reasonable size. A related limitation is that PSM is not possible unless there is sufficient overlap in the propensity score between the treatment and comparison groups.

A key PSM limitation is the need for a large, detailed data set of participants and nonparticipants.

However, the main limitation of PSM is the assumption that individual characteristics that are not accounted for are not correlated with outcomes or do not differ across the treatment and comparison groups. We can only match individuals based on the characteristics that are observed in the data. Thus, the critical assumption of matching methods is that there are no unobserved or unmeasured differences between the two groups that affect outcomes. This is a strong assumption in most settings.

For example, individuals who participate in a program are often motivated and, thus, likely to have better outcomes than the matched comparison group, even in absence of the intervention. If there are differences in unobserved characteristics that affect outcomes, then matching does not lead to comparable pairs. This means that the matching assumption fails and our estimate of the program impact will be biased.

Because of this assumption, PSM is not considered a strong research design and should only be used for impact evaluation when other experimental or quasi-experimental methods are unavailable or in combination with other methods (see section 6.5 for more details on combining methods).

6.4.3 Difference-in-difference

When none of the methods described above are feasible, it may be possible to obtain a rough estimate of the impact of the program by using a **difference-in-difference (DID)** evaluation design.

The key idea of DID is to combine two methods to improve inferences. Earlier in section 6.2, we discussed two methods that, when used alone, do not provide an estimate of the impact of the program. The first is the difference in outcomes before and after the program for those who participated in the financial capability program. The main concern with this difference is that any changes in outcomes could be the result of the intervention or another event that occurred in the time between the two data collections. The second difference compares the outcomes of participants and nonparticipants. Used alone, this is also problematic because these groups could differ in unobservable characteristics that may drive the differences in outcomes.

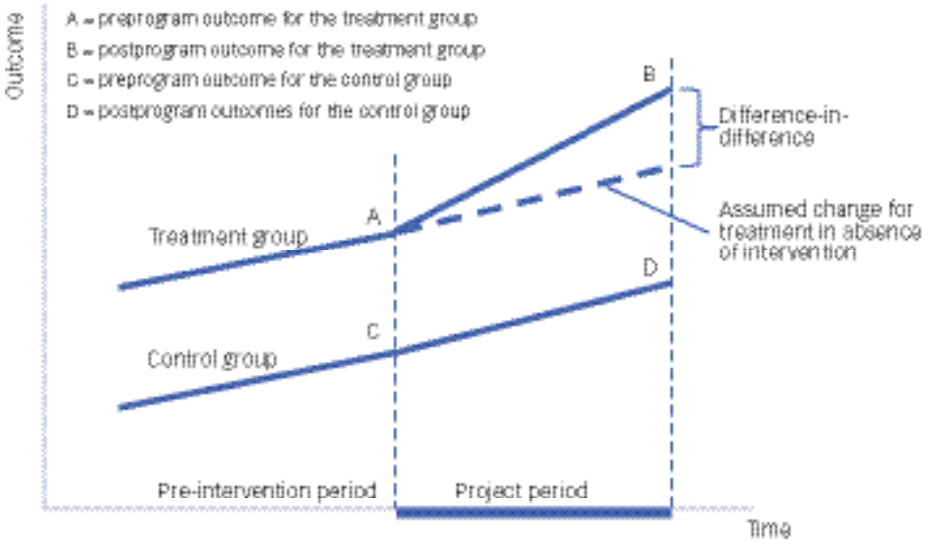
However, by combining both of the simple differences into a DID framework, we can address some of these concerns. The DID method gives the change in the outcome of the treatment group over time, relative to the change in the outcome of the comparison group. Here is one way of looking at the broad idea: In addition to the difference in outcomes among participants, we can also make use of the pre/post difference among nonparticipants. Just like the participants, the nonparticipants are subject to the concurrent events. But unlike the participants, the nonparticipants are not affected by the financial capability intervention. Thus, by calculating the difference between the pre/post differences in outcomes of the participants and nonparticipants, we can isolate the effect of the intervention over time.

DID gives the change in the treatment group outcome over time, relative to the change in the comparison group.

For example, consider a financial education workshop designed to encourage participants to sign up for a bank account. If we can measure the difference in bank usage for the treatment and control groups, both before and after the intervention, then the DID method can give us an estimate of the impact of the workshop on bank usage.

Figure 6.4 shows how we interpret the difference in the changes across the two groups in the DID analysis. Here we assume that participants in the treatment group have higher outcomes even before the intervention. This could be because they are motivated or because of some other characteristics. The DID estimate of the program impact is shown at the top on the right.

FIGURE 6.4 GRAPHICAL ILLUSTRATION OF THE DID DESIGN



DID's key assumption is that, in the absence of the treatment, the **change** in outcomes for the "treatment" group would have been similar to the **change** in outcomes for the comparison group. This is often called the "parallel trends" assumption and allows us to calculate the counterfactual for impact evaluation. The two groups do not need to be exactly comparable at the baseline. We merely require that their trajectory over time would have been similar. Translating this assumption to figure 6.4, we have assumed that the slope of the dotted line is the same as the slope of the control group line during the "intervention period."

BOX 6.11 RTF IN PRACTICE: APPLYING DID TO THE EVALUATION OF A SAVINGS PROGRAM, NIGERIA

This RTF project explores how new media and "learning-by-doing" can encourage financially unsophisticated consumers to open and maintain savings accounts. A large Nigerian bank launched a nationwide savings promotion called "I-Save I-Win" (ISIS). It featured a large number of heavily publicized lottery prizes for those who opened or maintained savings account and maintained savings balances above various threshold amounts for 90 days. To study how the program influences the savings activity of existing account-holders, the researchers conducted a DID analysis: they compared savings account balances before and after each phase of the ISIW promotion. To control for general trends in savings activity over the same time period, which may not be due to the savings promotion, they use data from a second bank that did not have a savings promotion to create a comparison sample.

However, this is not a trivial assumption. For instance, the participants may not only receive financial literacy training but also benefit from a business management workshop that was conducted at the same time. If the nonparticipants did not join the workshop, the DID will identify the impact of **both** workshops on outcomes, not just the impact of the literacy training. In fact, all events that affect only the participants will be bundled in the "program effect" in the DID methodology.

Said differently, the DID method works if the two groups have the same experience over time, with the sole exception that one group participated in the financial capability program. Other events that affect the outcomes differentially will create problems for this evaluation method.

Other events that affect outcomes will create problems for DID.

PROGRAM CONDITIONS FOR SELECTING THIS EVALUATION APPROACH

DID can be used if:

- **There is reasonable certainty that the participants and nonparticipants only differ in their exposure to the financial capability program.** If either group has been affected by some other event, the DID method will lead to a biased result.
- **There are at least two instances of data collection for participants and nonparticipants: before and after the intervention.** Additional preprogram data are not critical, but they are useful to implement robustness checks (see below).
- **The same individuals can be followed over time**; while this condition is not necessary, it is the preferred situation.

Recall that the purpose of taking the difference over time is to account for all characteristics (both observed and unobserved) that do not change over time. This method works best if we observe the same individual twice: We can immediately eliminate the time-invariant factors by taking the difference between the two points in time for each individual. This particular method is often called a **fixed-effects DID**. However, data on different sets of people before and after the intervention is also useful. This approach is called a **repeated cross-section DID**.

IMPLEMENTATION

In its simplest form, DID can be implemented with a comparison of means. Table 6.1 shows an example using the following steps:

1. Calculate the pre/post difference in the outcome for the program participants (B–A).
2. Calculate the pre/post difference in the outcome for the nonparticipants (D–C)

TABLE 6.1 IMPLEMENTING DIFFERENCE-IN-DIFFERENCES

	BEFORE	AFTER	DIFFERENCE
Participants	A = 3	B = 5	(B–A) = 2
Nonparticipants	C = 1	D = 2	(D–C) = 1
Difference			(B–A) – (D–C) = 1

3. Calculate the difference of the differences by subtracting the difference for nonparticipants from the difference of the participants (B–A) – (D–C).

The order of the differencing does not matter, so you can take the difference in participants and nonparticipants—first for before the program and then for after the program—and take the difference in these differences to get the program effect. These three steps can also be implemented in a regression framework that allows you to control for individual characteristics. The regression approach also makes the generalization of the DID to multiple years straightforward, which is useful if program participants can be followed over multiple years. (For more details, see appendix B.)

Although it is difficult to directly test the assumptions made in DID, there are two ways to provide indirect checks. The first is to use any available data predating the intervention to show that, before the intervention, treatment and comparison groups showed the same trends in outcomes. A second alternative is known as a **falsification check**—implementing a DID methodology for outcomes that should logically **not** be affected by the program. If the assumptions for the DID method are correct, then these DID estimates should be close to zero; otherwise, this suggests that larger trends may be at work and the assumptions are false.

LIMITATIONS

The key limitation for the DID method is that it does not use randomization, but instead relies on the assumption that the change in the participant outcomes would have been the same as the change for the nonparticipants, in the absence of the intervention. That is, we have to assume that there are no other differential time trends between participants and nonparticipants, because any difference in the changes between the two groups is attributed to the program. Like the PSM method described above, the DID method makes strong assumptions about the treatment and control group and is, thus, only suggested in combination with other methods or in situations where other methods are not possible.

BOX 6.12 RTF IN PRACTICE: SCHOOL-BASED FINANCIAL EDUCATION IN BRAZIL

In the case of the Brazilian school-based intervention, the evaluation team used matching, randomization, and difference-in-differences. Schools were recruited into the study from six districts. Of the 900 schools that voluntarily indicated willingness to participate, schools were ranked by their propensity scores and assigned to pairs. One school in each pair was randomly selected to receive the textbook and training, while the remaining school was scheduled to receive the textbook after the pilot program was completed. Treatment effects were then estimated by comparing the average changes in student performance from baseline to follow-up survey, between treatment and control schools. Both the treatment and control schools improved over time (as students continued to learn other relevant skills in the regular curriculum), but students in treatment schools did significantly better on measures of both objective financial knowledge and other measures, such as autonomy.

6.5 COMBINING IMPACT EVALUATION METHODS

Although all the impact evaluation methods discussed above have some limitations, combining methods can help offset the limitations of any one method and create a more robust counterfactual. In practice, most of the methods described are used in combination.

Indeed, whenever baseline and follow-up data are available, the DID method may be applied in combination with any of the other methods reported. As noted earlier, the PSM method does not generally create the most robust counterfactual. While matching reduces selection on observables, the issue of selection on unobservables remains unresolved. **PSM with DID** (i.e., comparing changes from baseline between a matched treatment and comparison group) strengthens the counterfactual by also controlling for time-invariant differences between the treatment and comparison group (whether observed or unobserved). Similarly, **RDD with DID** can be implemented by comparing changes in outcomes over time for units just above and below the program eligibility thresholds. Even when treatment and comparison groups are randomized, the collection of baseline data and use of **RCT with DID** is useful. The use of panel data leads to more statistically precise estimates of treatment effects (because repeated measurements tend to be correlated). It can also be helpful to separately estimate time effects versus program effects and to understand background changes that are simultaneously occurring.

In general, wherever possible, "worst case" scenarios should be built in. Combining different research designs can give you such a fallback position if one of the designs fails.

6.6 PRACTICAL AND LOGISTICAL CHALLENGES TO IMPACT EVALUATION

Often, even if a design is theoretically possible, practical challenges to implementation remain. Here, we list some those practical and logistical challenges.

6.6.1 Compromises implementing and maintaining randomization

There are several practical challenges related to any randomized evaluation, regardless of whether the intervention involves financial capability. Even after some form of randomization is agreed-upon by major stakeholders, **equity** often remains a significant concern for evaluators, implementers, funders, politicians, and beneficiaries. Specifically, when the intervention is considered particularly helpful and only a fraction of the eligible population receives the intervention, or when the intervention is delayed to some of the population, there may be difficulties implementing the randomization.

Equity is a significant randomization concern for all evaluators and all stakeholders.

These concerns can have strong practical implications for evaluation. While a program's major policy makers may see the value of randomization, an RCT may be seen to conflict with certain program goals or objectives. There may be vocal opposition to the study design, leading to a lack or withdrawal of support from key stakeholders. External funders may wish to implement programs where benefits are most likely to be largest, while politicians may wish to do so where votes are most valuable. At a more grassroots level, local authorities, implementers, and the eligible population itself may perceive the randomization negatively. A lack of support may reduce the efficacy and reach of the program and could reduce its impacts.

Also, there are significant practical difficulties in performing the randomization and ensuring that it is maintained. Very commonly, lists of participants or clients used for randomization may not be up-to-date—especially if operations are largely decentralized—and implementers may find that realities on the ground differ from their instructions. Implementers in existing programs that are already highly burdened may find the additional requirements of maintaining a randomization difficult to comply with. Moreover, equity or other concerns among implementers may lead them to deliberately ignore randomization requirements and provide services to all eligible beneficiaries.

Mitigating the risks from misunderstandings or misgivings about randomization requires a strong communication strategy that will clarify the meaning of—and build support for—the evaluation design to all relevant stakeholders. In particular, implementers must be apprised and supportive of the requirements of randomization. If treatment is blind, information sharing should be carefully managed, and if not blind, communication about the goals and objectives of evaluation may also need to encompass the control group.

Mitigating randomization risks requires a strong communications strategy.

6.6.2 Difficulty with eligibility rules and protocols

Some quasi-experimental research designs also have logistical challenges. For example, RDD requires that participants and nonparticipants be separated by a relatively clear rule. Evaluators can estimate the project's impact by comparing outcomes of individuals just below and just above the cutoff. But sometimes rules are not respected in practice. Administrators and officials may give a benefit to an individual who should have been ineligible by a small margin. If the rule is publicly known, these individuals may also misrepresent their own status to participate, which can lead to a failure of the RDD assumptions. In the RDD setting, it is advisable to keep the assignment rule nontransparent to prevent the manipulation of the treatment and control groups by stakeholder.

6.6.3 Maintaining control group integrity

One possibility to always consider for an RCT is whether the behavior of the control group may change in response to the presence of the randomization itself. One type of reactivity is when the control group perceives itself to be in competition with the treatment group. In a school-based intervention, for instance, teachers who may not be randomized into receiving supplemental financial education support material may try to actively compete in other ways.

Other forms of reactivity may not involve active competition but can still be problematic for a counterfactual: Individuals who are excluded from a financial training intervention may simply become aware of such programs and decide to obtain training elsewhere.

Finally, being denied an intervention may result in negative behavior that compromises other parts of the evaluation, such as attrition; even if randomization is staggered, those who are randomized to not receive initial treatment may withdraw from a program entirely.

A related problem is the Hawthorne effect. This refers to the tendency of subjects to modify their behavior in response to being evaluated (rather than because of the effects of the specific intervention). The original story about Hawthorne effects

Hawthorne effect: the tendency of subjects to modify their behavior in response to being evaluated

describes workers at an auto plant improving their productivity—albeit only in the short-term—in response to a study. This effect may be present in participants of financial capability interventions, especially when repeated surveys make certain behaviors more salient.

For example, frequent recording of financial transactions may help subjects learn to track their finances independent of any intervention. This can also affect implementers in situations where there are generally recognized issues of motivation and provider performance. In health care, for instance, observational studies of physicians in developing countries have been shown to increase the quality of the services provided. Similarly, in the financial education context, being observed may lead teachers to exert more effort when teaching. Alternatively, certain providers may choose to adhere more scrupulously to protocols or even refrain from taking informal payments while in a study.

BOX 6.13 RTF IN PRACTICE: EXAMPLE OF THE HAWTHORNE EFFECT

In the RTF soap opera financial capability storyline implemented in South Africa, researchers were concerned about the Hawthorne effect—that program participants would modify their behavior in response to being evaluated.

To avoid this situation, soap opera viewers in the treatment and control groups were both given encouragement. Both groups were asked to watch different soap operas, with only the treatment group being assigned to watch *Scandal!*, the intervention. Both groups were then interviewed with the same frequency and asked similar questions about the content of their respective soap operas.

KEY POINTS

For program management, policy makers, funders, and other stakeholders, impact evaluations provide critical input on whether financial capability programs deliver on what they promise. But determining whether the program or something else is responsible for the impacts we see is very hard to do. It is fairly easy to see effects and argue that a financial capability program is responsible for those effects, but measuring these effects is another matter.

Depending on the situation, there are a series of quantitative approaches that evaluators can use to help tease out the impact of financial capability programs from other factors. Experimental approaches—like RCTs and encouragement designs offer the most robust approaches for demonstrating impact from an intervention, but they are sometimes not feasible given specific situational conditions. Other quasi-experimental approaches—like RDD, PSM, and DID approaches are not as robust but are

still valuable tools for measuring and showing impact. All of them are better than before-and-after designs and comparisons of outcomes between voluntary participants and nonparticipants, neither of which by itself is good enough to make causal arguments about program impacts. Table 6.2 summarizes the implementation timeline (i.e., whether the methods are used before, prospective, or after, retrospective, the program is implemented) and the strength of the methodology in terms of estimating causal effects described in this chapter.

TABLE 6.2 OVERVIEW OF IMPACT EVALUATION METHODS

EVALUATION METHODOLOGY	IMPLEMENTATION TIMELINE	CHANGE CAN BE ATTRIBUTED TO INTERVENTION	STRENGTH OF METHODOLOGY
Randomized Control Trial	Prospective	Yes	****
Encouragement design	Prospective or retrospective	Yes	***
Regression discontinuity	Retrospective	Yes	***
Propensity score matching	Retrospective	Yes	**
Difference-in-difference	Retrospective	Maybe	*
Pre/post comparison	Retrospective	No	—
Comparing participants and nonparticipants	Retrospective	No	—

FURTHER READING

General

Bamberger, M. 2009. "Institutionalizing Impact Evaluation within the Framework of a Monitoring and Evaluation System." Washington, DC: World Bank,

Gertler, Paul J., Sebastian Martinez, Patrick Premand, Laura B. Rawlings, Christel M. J. Vermeersch. 2011. Impact Evaluation in Practice, Washington, DC: World Bank.

Technical

Agarwal, S., G. Amromin, I. Ben-David, S. Chomsisengphet, and D. D. Evanoff. 2010. "Learning to Cope: Voluntary Financial Education and Loan Performance During a Housing Crisis," *American Economic Review: Papers and Proceedings*, vol. 100, 495–500.

Angrist, J. D., and A. B. Krueger. 2001. "Instrumental Variables and the Search for Identification: From Supply and Demand to Natural Experiments," *Journal of Economic Perspectives* 15 (4): 69–85.

Blundell, R., and M. Costa Dias. 2002. "Alternative Approaches to Evaluation en Empirical Microeconomics," *Portuguese Economic Journal* 1(2): 91–115.

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

Imbens, G. W., and J. M. Wooldridge. 2009. "Recent Developments in the Econometrics of Program Evaluation," *Journal of Economic Literature* 47 (1): 5–86.

Jacob, B. A. and Lefgren, L. 2004. "Remedial Education and Student Achievement: A Regression-Discontinuity Analysis," *Review of Economics and Statistics* 86 (1): 226–44.

Jalan, J., and M. Ravallion. 2003. "Estimating the Benefit Incidence of an Antipoverty Program by Propensity-Score Matching," *Journal of Business and Economic Statistics* 21 (1): 19–30.

Meyer, B. D. 1995. "Natural and Quasi-Experiments in Economics," *Journal of Business & Economic Statistics* 13 (2): 151–61.

Rosenbaum, P., and D. Rubin. 1983. "The Central Role of the Propensity Score in Observational Studies of Causal Effects," *Biometrika* 70 (1): 41–55.

CHAPTER 7

Putting it all together

While so far we've certainly discussed monitoring and evaluation as events that often occur in conjunction with one another, for the most part we have treated those discussions in chapter 4 (Monitoring), chapter 5 (Process Evaluations) and chapter 6 (Impact Evaluations) as discrete events for the sake of making it clear what goes into each type of evaluation. But in the real world, monitoring and the various forms of evaluation typically take place together as part of a comprehensive evaluation of a financial capability program, although they generally take place during different stages of the program's life cycle.

In this chapter, we take this comprehensive evaluation perspective, discussing what we refer to as "mixed-method" approaches that integrate the evaluation methods we have already discussed to provide a more comprehensive picture of evaluation.

7.1 WHY IS A COMPREHENSIVE EVALUATION SO IMPORTANT?

In the previous chapters, we discussed why it is critical to properly design an impact evaluation, but it is equally critical to keep the role of an impact evaluation in perspective. An impact evaluation is fundamental to being able to credibly attribute effects to a program, which is what program implementers and other stakeholders are most interested in. Based on the program's theory of change, we expect that implemented financial capability programs will be effective in accomplishing the goals and objectives set out; the bottom line is to improve the financial capabilities of the populations served with the services provided.

Comprehensive evaluations let us look into the "black box" of overall program effect and help interpret findings.

But we also need to be able to look into the "black box" of the overall program effect to answer more basic questions about its theory of change and to help interpret the findings. The term "black box" often refers to modeling results, where there is a sense that researchers conduct an analysis and results are generated—results that are opaque because we may not understand what went into the model to help us understand what came out of it. Here, we use "black box" to refer to program effects, which can be equally opaque without proper context or interpretation. For example:

- **If a program succeeded**, what were the underlying mechanisms that led to behavior change? What can this experience tell us about the most relevant barriers and constraints to financial capability to address in the target population, or other similar groups? What have we been able to learn about designing better programs in the future?
- **If a program did not meet its objectives**, why did it fail or fall short? Was there a fundamental conceptual problem (faulty program logic)? Were the participants inappropriate (faulty targeting)? Was it simply poorly executed (faulty implementation)?

Answers to these kinds of questions can help program implementers and stakeholders better understand what effects they are seeing and can make an evaluation valuable even if the program itself is not considered a success.

For example, consider a financial training program at a rural bank aimed at increasing savings that is randomly implemented in some branches but not others. The program seems promising, and enrolls large numbers of participants in treatment branches at the start. However, at the end of a year, the impact evaluation shows no results: Participants in the treatment group have saving and credit behavior that is no different than that of the control group who didn't receive the program.

Understanding the impact of program can be improved by a more comprehensive evaluation that puts all the pieces together.

When interviewed, bank officers suggest some potential reasons for why the impact evaluation didn't show the expected outcome. For example, many participants may have dropped out early because the program was too long, the program content may not have focused sufficiently on savings, or the teaching may not have been of sufficiently high quality. Each of these reasons, moreover, suggests a different response to improving the program in the future or to improving programs like it in the future: Shorten the program, change the curriculum, and/or raise teaching standards. Thus, our understanding of the impact of a program—whether it went as expected or didn't—can be improved by a more comprehensive evaluation that puts all the pieces together.

7.2 HOW DO WE CONDUCT COMPREHENSIVE EVALUATIONS?

As the above questions and example indicate, having more information about how a financial capability program is conducted is necessary to understand what happened. But those kinds of information don't always come from the impact evaluation alone; some of that information will come from understanding how and when services were delivered and how the program was implemented in the field. In other words, an

impact evaluation needs to be combined with the high-level knowledge of program implementation gained through monitoring and process evaluations.

The example also argues for using the different types of data—qualitative and quantitative—that are a part of the different evaluation types (in the next chapter on Data Collection, we discuss the different types of data in detail). Such "mixed methods" naturally arise in an evaluation strategy that combines impact evaluation, process evaluation, and analysis based on monitoring data. Combining different sources of information and evaluation methods allows us to assess both the "what" and the "why" of program evaluation: Impact evaluation methods help to estimate the magnitude of impacts, and process evaluation methods help shed light on the underlying causal mechanisms and processes.

Combining different sources of information and evaluation methods allow us to assess both the "what" and the "why" of program evaluation.

Such a mixed-methods approach recognizes the complementary strengths and scope of all the evaluation methods discussed previously in this Toolkit. Impact evaluation allows evaluators to arrive at statistically meaningful causal estimates of a program's effects, which can make the findings of the evaluation ultimately more useful to policy makers and the wider development community. For instance, finding that a program is "positively associated" with improved financial behavior is useful, but less actionable than evidence that the program has had a 5 percent impact on a particular target measure, which provides a tangible and concrete benchmark for comparison against other programs.

While qualitative methods cannot accurately estimate the **magnitude** of a program's effect on outcomes—even when comparison groups are used (such as having focus groups with nonbeneficiaries)—but qualitative methods can help overcome the limitations of purely quantitative approaches by providing a less-structured, open-ended, and flexible method of inquiry. Such inquiry allows for the discovery of effects on the beneficiaries, program staff, and other stakeholders that were unforeseen by evaluators. It can provide a useful alternative avenue for further exploration of research questions when extended quantitative analysis is limited by budget, time, or data constraints. Qualitative research methods that are participatory in nature can also increase the acceptability and utilization of findings from quantitative inquiries.

7.3 WHEN IS A COMPREHENSIVE APPROACH MOST CRITICAL?

For all the reasons above, conducting a more comprehensive evaluation that integrates both impact and process evaluations and the data they yield makes sense. Doing so can obviously go beyond the resources allocated, but there are certain situations where it makes the most sense to strive to do so. We list below some of those situations.

BOX 7.1 RTF IN PRACTICE: COMPREHENSIVE EVALUATION OF FINANCIAL TRAINING PROGRAM IN INDIA

One RTF pilot project involves a financial capability program in Bihar, India, that is seeking to change individuals' financial behavior and knowledge. Those living in rural villages in India typically have to travel long distances to access a bank to conduct financial transactions, a situation that creates disincentives to using a bank. To address this concern, EKO India Financial Services has provided a "doorstep banking" program in rural villages—a program that allows people to conduct financial transactions close to their homes.

In addition to gaining access to the program itself, a sample of program participants also received a financial training course on the importance and benefits of saving. The program evaluator conducted a randomized control trial (RCT) to evaluate the impact of the training course. Randomization was at the village level; within the treatment villages, a single family member from a fraction of the families receives financial training, while the remaining families have multiple members who receive financial training.

To disentangle how attitudes and behaviors were shaped by the financial capability intervention, the project team relied on a comprehensive evaluation. Beyond the RCT that forms the impact evaluation, the project team conducted a process evaluation using in-person interviews and focus groups. These interviews were conducted with program staff and program participants immediately after the intervention was delivered. The questions addressed a number of areas, such as **service delivery** (e.g., location and timing of training and feasibility of program approaches used), **resources** (e.g., whether resources were adequate and how to reduce resources without compromising program quality), **service use** (e.g., how participants were recruited and which participants took up the program), and **program participant experiences** (e.g., what the motivation was for participating and opinions about the length of training and the content of the course).

7.3.1 Conducting the exploratory (or pilot) phase of an evaluation

In this situation, quantitative impact estimates are important as key resource decisions are being made. As the same time, qualitative data collection can be used to develop hypotheses, research questions, and survey instruments, as well as to define indicators for quantitative evaluation. Often, qualitative data from pilot studies are used to design interventions themselves. We discuss the advantages of using qualitative data for exploratory research in further detail in chapter 8.

In our bank example from above, after a pilot program, decision makers will want to know if scaling up is warranted, especially considering the opportunity costs of allocating capital to the program. This means that evaluation results need to address the realized as well as potential effect sizes from the program, and associated costs, so a quantitative impact evaluation is needed to capture these numbers in a way that allows such computations. At the same time, those implementing the program could have conducted focus groups with respondents prior to the beginning of the program

being implemented. Doing so may have helped the bank to identify the most common barriers to savings in their customer base and, thus, helped in designing the program in a way that would hopefully get around those barriers. Process evaluation and qualitative research may also help decision makers understand important ways in which a scale-up might differ from the pilot program experiences.

7.3.2 Following up after impact evaluation and understanding differences in program effects

The type of quantitative analysis we have described in our impact evaluation chapters can establish **average** program effects by comparing a large group of participants to another comparison group. But quite often, evaluators are interested in going beyond the average and looking at potential differences in effects for subgroups (for instance, women versus men, the more educated versus the less educated). Ideally, we would wish to perform a quantitative analysis on several different groups, drawing large samples for each of the subgroups of interest. However, it is often logistically or financially infeasible to gather enough participants in each subgroup to be statistically meaningful; as a result, many evaluations are not "powered" to do detailed subgroup analysis. More discussion of heterogeneous effects and the issue of having enough statistical power in analyses can be found in the chapters on impact evaluation design (chapter 6) and data analysis (chapter 10).

The quantitative analysis of survey data can identify different effects among different groups, and in-depth case studies can help understand and describe these groups in greater detail.

For instance, we may be interested in comparing program effects for men and women, particularly widows, but the budget may not permit an evaluation of men and women separately. Furthermore, it may be difficult to recruit a large sample of widows. In such situations, it can be useful to draw on qualitative data collection methods, such as interviews or focus groups, because such methods can help evaluators understand why certain people benefited much more or much less than the average person in terms of whatever program effect found in the quantitative analysis.

Similarly, evaluators could analyze quantitative data from a survey to identify different or **heterogeneous** effects among different groups (e.g., those respondents who budget annually, those who budget monthly, those who budget irregularly, and those who do not budget at all), while in-depth case studies (a qualitative method) can help describe and understand these groups in greater detail. The survey will provide breadth, while the case studies will provide depth.

7.3.3 Increasing the validity of evaluation findings

Sometimes the findings from the impact evaluation may be either unexpected or controversial, perhaps so much so that policy makers and stakeholders may doubt

their validity. One way to show the validity of the findings is to use a technique known as "triangulation."

Triangulation combines data sources to address a particular question, thus providing more weight to findings.

Triangulation involves combining various data sources to address a particular research or evaluation question, thus providing more weight to the findings. This can, of course, be done within quantitative or qualitative data. In the former, triangulation is sometimes achieved by analyzing two or more different surveys, thus showing that the unexpected or controversial result is not just an anomaly in the current evaluation. In the latter case, triangulation may consist of conducting focus groups alongside participant observation or key-informant interviews—again, confirming the result with multiple sources. Within the context of a mixed-methods approach, triangulation may involve using interviews to confirm the findings from a survey to confirm or validate findings.

7.3.4 Case studies of small or very sensitive populations

Surveys can provide incredibly valuable support in evaluations, but survey data may not allow for the proper examination of certain marginal markets or activities that may be important for programs and policies. In other cases, populations may be too small to warrant any kind of quantitative analysis but are still worth investigating.

Case studies provide an in-depth understanding of how a phenomenon functions and can answer question of "how" and "why" a phenomenon came to pass.

For instance, in our example, suppose it is difficult to interview elderly widows in this population using conventional survey methods, but this population is important because it is particularly vulnerable. A more appropriate way to obtain information may be to use interviews or very small focus groups. While the nature of the evidence gathered may not be comparable to conducting a large impact evaluation, this form of inquiry may be the best way to reach the population of interest and perhaps also to obtain more nuanced information, albeit more limited in generalizability.

Another way to do this is through case studies, which is an approach used to investigate a phenomenon in its real-life context. Case studies typically consist of the detailed examination of a particular individual, a small group, an institution (a household, a company), event, or setting, and are often conducted as part of broader research endeavors. Case studies allow researchers to gain an in-depth understanding of how a phenomenon functions, and can answer questions of "how" and "why" a phenomenon came to pass.

In many types of qualitative research, case studies are often used as **illustrations** of common phenomena or to examine contrasting phenomena. Such illustrative case studies are referred to by a number of names, including collective case studies, multiple case studies, cross-case studies, comparative case studies, and contrasting case studies. When case studies are used as illustrations of common phenomena,

the ultimate goal is usually to improve the researcher's ability to understand and theorize about a broader context. Moreover, using multiple case studies makes findings more robust than a focus on a single case study.

In impact or progress evaluations, the evaluation of a financial capability program that provides services to many dispersed communities may select a small number of communities as case studies of how service provision occurs on the ground. Doing so allows for greater depth in examining the processes, stakeholders, and interactions in place, providing informative evidence on the "mechanics" of a financial capability program. This information, in turn, can be used in future decision making about the implementation of this or similar programs.

Similarly, a program evaluation can use case study households to provide an in-depth description of the way the program's services or transfers are used, shared, and perceived within that household. While these kinds of case studies cannot explain how **all** households interact with the program or how the program delivers services in **all** communities, they can offer informative insights into program issues that may be relevant more generally.

Data for case studies can gathered through various data gathering techniques, including qualitative and quantitative methods. The case study method is based on triangulation of information (as discussed above) from multiple sources of evidence, which typically include interviews and/or focus groups, reviews of materials and documents, and direct observation.

Properly conducting a mixed-methods evaluation can be challenging. Evaluators should be careful not to embark on conducting such an evaluation without a clear rationale for why this approach is necessary. You should also think carefully about the mix of skills and expertise in your evaluation team. If your team consists only of researchers with quantitative analysis skills, it may be difficult to successfully incorporate insights from qualitative data into your evaluation and interpretation of results, and vice-versa. It may also be important to ensure that experts with different methodological strengths communicate effectively and cooperate well. In other words, it is important not to take a casual stance to integrating these approaches. The use of mixed methods requires solid reasoning and careful planning; otherwise, it may just drain resources from your evaluation's limited budget, staff, and time frame.

The use of mixed methods requires solid reasoning and careful planning; otherwise, it may just drain resources.

Table 7.1 provides a checklist of some things to consider if want to conduct a mixed-methods evaluation.

TABLE 7.1 CHECKLIST FOR DECIDING ON A MIXED-METHODS EVALUATION DESIGN

Consider your primary evaluation questions and think about which methods best address each of them ▪ For example, treatment effect questions will be answered using quantitative data analysis methods, process questions will be answered using qualitative methods, and effects heterogeneity can be answered using either or both	X
Check whether results from the analysis of data leaves unanswered questions that could be explored using other methods ▪ For example, if your randomized controlled trial identifies extreme cases (such as a few people who were worse off after the intervention), you may want to add some qualitative data collection and analysis to gain insight into why this may have happened	X
Think about your qualitative data ▪ For example, can it be coded and analyzed quantitatively to provide different kinds of insights?	X
Make sure the evaluation team includes people with skills and expertise in each of the different methods you will use in your evaluation ▪ For example, does the team have people who can conduct surveys and people who can conduct focus groups?	X
Think about your budget ▪ For example, is the budget sufficient to do a mixed-methods evaluation without spreading resources so thinly that it does not add much depth to your inquiry?	X

KEY POINTS

While an impact evaluation is **necessary**, it is not **sufficient** to achieve a full understanding of how a program performed. Process evaluation and monitoring are also important parts of a comprehensive approach to evaluation, helping to answer the questions about "how" and "why," which ultimately make the overall evaluation's results useful for program decision making. Every program and environment has unique features and resources, often calling for different ways of putting a comprehensive evaluation together. Combining information about effectiveness with cost information can then support analyses to increase program efficiency, which we will discuss in chapter 11.

FURTHER READING

General

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

Berg, B. L. 2007. *Qualitative Research Methods for the Social Sciences*, Pearson, NY: Allyn & Bacon.

Rao, V., and M. Woolcock. 2003. "Integrating Qualitative and Quantitative Approaches in Program Evaluation," In F. J. Bourguignon and L. Pereira da Silva, eds., *The Impact of Economic*

Policies on Poverty and Income Distribution: Evaluation Techniques and Tools, New York: Oxford University Press, pp. 165–190.

Yin, R. 1994. *Case Study Research: Design and Methods* (2nd ed.), Thousand Oaks, CA: Sage Publishing.

Technical

Greene, J., and Caracelli, V. 1997. "Advances in Mixed-Method Evaluation: The Challenges and Benefits of Integrating Diverse Paradigms: New Directions for Evaluation" (Series: JB PE Single Issue (Program) Evaluation).

Sandelowski, M. 2000. "Combining Qualitative and Quantitative Sampling, Data Collection, and Analysis Techniques in Mixed-Method Studies," *Research in Nursing & Health* 23 (3): 246–55.

PART III

Collecting and analyzing M&E data for financial capability programs

CHAPTER 8

Data collection methods

So far, we've discussed monitoring and the types of evaluations—process and impact—that are typically used in tracking and assessing financial capability programs. But key to implementing a monitoring system and any type of evaluation is data collection and, then, ultimately data analysis. There are both qualitative and quantitative methods for collecting data; while the qualitative variety typically goes with process evaluations and the quantitative variety with impact evaluations, both methods can and do get used in both, as we discussed in the previous chapter on comprehensive evaluations.

In the previous chapters, we mentioned data collection methods as part of the discussion. Here, we go into more detail on the types of data collection methods and their uses in evaluating financial capability programs. There are also practical concerns in actually implementing any of the methods discussed here; we save that discussion for the next chapter.

8.1 QUALITATIVE DATA COLLECTION METHODS

What is the purpose of using qualitative data in monitoring and evaluation? Qualitative data complement quantitative data by providing a depth of contextual understanding and a level of detail that one can't get with quantitative data alone.

Qualitative data can shed light on the causal mechanisms of a program—helping understand how program effects came to pass (or failed to do so).

When do qualitative data get used? There are a number of points in an evaluation where qualitative data can be useful. First, they can be used to help develop the quantitative research materials, such as survey questionnaires—both in terms of the substance of the material on such questionnaires and in terms of question wording and syntax. For example, evaluators of a financial capability program may conduct a small number of focus groups (qualitative data) to examine how potential program beneficiaries talk about the issues the program addresses (such as savings, budgeting, and so forth), which can then help construct the surveys (quantitative data) using the most appropriate terminology.

Second, as part of a **formative process evaluation**, qualitative data can provide a detailed understanding of the development of the financial capability program as it

is being implemented. In particular, qualitative data can shed light on program implementation and operational issues, including questions about the most appropriate mode of delivery, the identification of the target population, and so forth. This most commonly involves research with the people responsible for designing and delivering program services, but it may also include program beneficiaries at this stage. The goal of gathering information at this stage is to improve the design and delivery of the program.

Third, qualitative data can also be a very valuable part of **summative research**, including both process and impact evaluations. Although qualitative data can't by themselves establish causality between the evaluated financial capability program and observed outcomes, they can help produce a richer and more informative picture of the program being evaluated and provide insights that can't be fully captured through surveys and other quantitative data.

They do so, in particular, by shedding light on the causal mechanisms of the program—by allowing us to ascertain **how** any observed program effects came to pass (or failed to do so). They can also help provide an understanding of the experiences of beneficiaries in the extreme ends of the distribution—that is, those who benefited much more or much less than the average. In this way, qualitative methods can fill interpretation gaps left when only quantitative methods are used, adding depth and texture to the inquiry. And ultimately, the insights produced through qualitative research can inform future program development and implementation.

At the highest level, qualitative data collection methods focus on gathering information on people's perceptions and experiences with the process of implementing a financial capability program and on the meaning they ascribe to particular events in that implementation. A number of qualitative data collection methods are typically used, each with their strengths and weaknesses. In the subsections that follow we describe the following techniques: in-depth interviews; semi-structured and cognitive interviews, focus groups, reviews of documents and materials, and audit studies or mystery shopping. These discussions go into some detail, indicating when you might want to use them. We discuss some of the practical matters of how data collection is done in the next chapter.

8.1.1 In-depth interviews

Evaluators typically use in-depth interviews to explore in detail the expectations, experiences, views, and feelings of the interviewee. They can be particularly useful for discussing personal details or sensitive issues that people may be hesitant to discuss in a group setting. In an in-depth interview, the role of the interviewer is to guide the respondent through the topics that need to be covered without

constraining the respondent to specific questions. They are also referred to as "depth interviews."

In-depth interviews are useful for discussing personal details or sensitive issues that people may be hesitant to discuss in a group setting.

In designing in-depth interviews, one typically starts with a full list of issues or topics to be covered in the interview and then groups them in a logical order. The next step is to generate from this full list what is known as a "topic guide," which the interviewer will use in the interview. This is a list of topics that can be used to aid the interviewer's memory in doing the interview.

8.1.2 Semi-structured and cognitive interviews

Unlike in-depth interviews, which use the topic guides shown in the extracts in box 8.1, semi-structured interviews use an interview protocol, prepared in advance, which specifies the wording and sequence of questions and provides prompts for the interviewer. Questions are "open-ended," meaning that respondents answer the question as they see fit, using their own words. While in some instances interviewers stick strictly to the protocol, most semi-structured interviews include probing on the basis of the interviewee's answers.

Cognitive interviews can help determine whether a question is fully understood, is being interpreted as intended by the interviewees, and is producing meaningful replies.

One situation when interviews are particularly valuable is in developing quantitative data collection materials, such as a survey instrument—a method known as cognitive interviews. In-depth interviews or focus groups (discussed below) may be carried out first to ascertain the content of the survey instrument and begin the process of developing the wording of the questions. Semi-structured interviews can be used in two ways in this context. Most simply, they can be used to determine possible response codes to questions; but they can also be used more creatively to test the wording of questions "cognitively"—to ascertain whether the question is fully understood, whether the interviewees are interpreting it as intended by the interviewers, and whether it produces meaningful replies from the interviewees. In this case, precoded replies may also be included to determine whether they are appropriate.

For program evaluation, semi-structured interviews are often the most useful and appropriate method, because they rely on an interview protocol designed in advance and used in the same way in every interview by every interviewer. As a result, semi-structured interviews increase the comparability of responses and enable clearer insights into particular issues of interest.

Semi-structured interviews rely on an interview protocol, which increases the comparability of responses.

Beyond program beneficiaries, program staff and others associated with the implementation of the program are also often interviewed, both to obtain factual information about the program and to capture their insights into issues of program design, delivery, and outcomes. These are typically referred to as **key informant interviews**.

While there is some overlap in many of the questions asked in such interviews of program beneficiaries and key informants, other questions are specific only to

BOX 8.1 TOPIC GUIDE EXTRACTS FOR IN-DEPTH INTERVIEWS

CitizensAdvice—the largest network of independent advice agencies in the United Kingdom that provide free advice, largely to low-income individuals—commissioned the Bristol Personal Finance Research Centre to conduct an evaluation of a U.K. evaluation of a financial education initiative with low-income tenants of a big social housing provider. One part of that evaluation involved in-depth interviews with staff members who helped deliver the program and with users of the program. Extracts of the topic guides are given below.

Extract of Topic Guide on Project Development (Interviews with Staff Members)

1. Describe your role in developing and selling the project. Probe:
   - Does the project build on an existing service or is it a completely new service?
   - Any involvement in engaging partners (e.g., community centers, credit unions) or in setting up sessions, venues, etc.?
   - What has been your involvement in developing the workshop materials (involvement of others)?
2. How easy or difficult has it been to get the project off the ground? Probe:
   - Compared with own expectations?
   - Compared with your experiences in setting up similar projects?
3. What have been the main challenges in getting started? Probe:
   - Engaging partners
   - Arranging sessions, venues, etc.
   - Reaching target groups
   - Project funding
4. How have you tried to overcome these challenges? How successful have you been? Probe:
   - Any help/advice? From who? How useful?
     - If no, would this have been useful? Possible providers?
     - Check if any contact with other projects
   - Drawn on lessons learned from similar projects?
   - Impact on delivery of project and meeting targets?
5. Have any other issues come to light since setting the project up?

Extract of Topic Guide on General Experience of the Workshop (Interviews with Program Users)

1. How did you find out about the workshop?
2. What were the main reasons you decided to attend the workshop?
   - What did you expect before you attended? How did you feel about it?
   - How different was it to any expectations you had?
3. What do you think was good about the workshop you attended? What did you get out of it?
   - Probe for location, length, format, content, facilitator, etc.
4. What, if anything, do you think could have been done better?
5. Was there any other information or advice you would have liked that was not covered in the workshop?
6. Have you discussed the information you received in the workshops with friends or family?
7. Have you recommended the workshop to friends or family members?
8. Would you be interested in attending similar workshops again in the future?

BOX 8.2 USING COGNITIVE INTERVIEWING IN THE DEVELOPMENTAL PHASE OF AN RTF PROGRAM

In the developmental phase of an RTF survey of financial capability, focus groups were used to develop the interview guide, with the wording of potential questions selected from previous surveys. But since these previous surveys had almost all been carried out in high-income countries, there were concerns that the questions would not be appropriate in a low- or middle-income country setting, particularly with the very poorest people within that setting. Some of the questions were open-ended ones, requiring precodes; others were designed to test ways of capturing replies, such as scales determining strength of agreement.

Semi-structured cognitive interviews were conducted to get feedback on question wording to address these concerns. All interviews were audio-recorded for subsequent analysis, and the interviewers completed a detailed feedback form during each interview. In the first round of 120 interviews, the following information was recorded for each question (if it applied):

- Question not understood and why
- Question not appropriate for respondents' circumstances and why
- Other comments, e.g., inconsistencies in replies across questions; questions that did not accurately capture respondent's level of financial capability
- Questions difficult for respondent to answer and why
- Other comments, e.g., inconsistencies in replies across questions; questions that did not accurately capture respondent's level of financial capability
- Suggestions for rewording question to improve understanding
- Verbatim responses to open-ended questions

We illustrate two examples of how the cognitive interviews highlighted potential issues with the survey questionnaire. First, for one question—"When I get my income I set priorities for things I must buy or pay first"—researchers were testing both whether the question was difficult to answer and a four-point agree/disagree scale respondents were supposed to use. The cognitive interviews revealed that a respondent had an issue in establishing priorities to pay and priorities to buy because he treated them as two different things. The interviewer asked him to consider these as one (bills and purchases) to give an answer. The respondent also found the scale difficult to use.

As another example, researchers were testing to see if a question was understood—"What are your main priorities"—and discovered that a respondent felt his priority was finding a job, which meant that the question was not totally clear because it referred to priorities about expenses.

A second round of cognitive interviews was then conducted with a slightly different focus and a revised questionnaire to take account of feedback from the first round. Feedback from the interviews and focus groups had a large impact in the development of the survey instrument. For example, the two questions in the above example ultimately became:

- When you receive money, do you plan how it will be used? Yes/No *(Interviewer instruction: if too little money to plan, code "No")*
- *Ask if yes at 12.* Do you always plan how the money you receive will be used or only do it sometimes? Always/Sometimes
- *Ask if yes at 1.* Do you plan exactly how you will use the money or only make a rough plan? Exactly/Rough plan

program beneficiaries or key informants. Table 8.1 provides examples of evaluation questions that can be asked for both types of stakeholders. Many of these can also be "topics" for discussion in in-depth interviews. It is important to note that there are other stakeholders (and nonbeneficiaries) that are often useful to interview. When deciding who to include as interviewees, it may help to make a list of possible types of interviewees and what information you hope to obtain from them. This will, in turn, help you develop the most appropriate interview protocol for each particular type of interview.

8.1.3 Focus groups

Focus groups elicit a person's own views in the context of those of others and allow that person to contribute to the views of others.

Focus groups (sometimes referred to as group discussions) involve the moderated discussion of a particular topic or topics with a small group of people (usually between 6 and 10). They are not designed to arrive at any kind of consensus nor be a decision-making forum. They are also not group interviews, where the moderator poses a question to the group and each respondent gives his/her view without an attempt to stimulate discussion among the group. Rather, focus groups are intended to elicit a person's own views on the topic of discussion in a context in which the person can consider the views and perceptions of other people and contribute to them. They also enable the researcher to observe the interactions between participants and the collective creation of meaning.

More specifically, focus groups can be cost-effective, can be used to obtain information from transient populations (e.g., prisoners, migrants, hospital patients), and can

TABLE 8.1 SAMPLE INTERVIEW QUESTION SET FOR PROGRAM BENEFICIARIES AND KEY INFORMANTS

QUESTIONS FOR PROGRAM BENEFICIARIES	QUESTIONS FOR KEY INFORMANTS
What were your expectations of the program?	
How did you decide to participate in the program?	What was the motivation for the development of the program?
What does your participation in the program consist of?	What is the nature of your involvement in the program?
What are your views of the way in which the program was delivered?	
What do you think are the main strengths or benefits of the program?	
What do you think are the main weaknesses of the program?	
Do you think the program has led to changes in your life, your family's life or the community? Why/Why not?	Do you think the program has led to changes in the lives of participants and their families, and/or the community? Why/Why not?
What kinds of changes would you suggest to improve the program?	

Note: Key informants are program staff, staff of institutions where program is implemented, local authorities, etc.

enable researchers to spot areas of shared or diverse experience in one fell swoop. Focus groups are also useful in other ways. For example, by their group nature, focus groups may raise new questions or point to aspects of the program being evaluated, which individuals in one-on-one interviews may not have considered before. Moreover, they can be especially useful when informants are too intimidated or otherwise unwilling to have one-on-one interactions with researchers. In fact, in some instances, researchers may find that focus groups are necessary before respondents agree to meet on a one-on-one basis for an individual interview. Then again, focus groups may be unsuitable to explore certain socially sensitive or highly personal issues. As a result, detailed questions about participants' lives and financial situations (such as employment and income) are usually kept to a minimum in focus groups. However, it is typically acceptable to discuss people's experiences and views on issues relating to personal finances in focus groups.

Given that there are similarities between interviews and focus groups, when should you use one versus the other? It is important to note that focus groups produce similar but distinct types of data. Focus group data reflect meanings, views, and opinions negotiated in the context of a group discussion. That means that data from an interview may differ from, and even contradict, that obtained from focus groups.

Focus group data are collected in the context of a group discussion and may differ from or contradict data collected in interviews.

In other settings, the program context will dictate the appropriate choices. For example, in a school-based program, it may make sense to conduct interviews with teachers to assess their general opinion of the materials, how easy they are to use, how clear they are, and how appropriate they are to the teaching environment. If time and resources allow, you may also want to conduct focus groups with students, who can tell you their views on whether the teacher's delivery has improved, lessons are more engaging, and the material covered in class is clear. Insights from both methods can help inform decisions on whether to retain or adjust the teaching materials.

Focus groups can be used at various different stages in an evaluation and for quite different purposes. Like interviews, focus groups can be used as part of a formative process evaluation to provide insight; for example, in a financial capability program, they can be used to explore content and delivery. And they can be used to inform the design of the content of a survey questionnaire that will be used in a summative evaluation involving either a process or an impact evaluation, or both.

In a summative evaluation, they can be used to provide a more detailed understanding of key findings. For example, a survey may show that users found a financial capability program very helpful, yet it had no impact on their behavior. Focus groups, like interviews, might help to explain why this is the case.

As was also true for interview preparation, a key step in focus group development is producing a full list of the things that should be covered in the discussion. Data

BOX 8.3 USING FOCUS GROUPS IN RTF PROGRAMS

We highlight developmental work for the RTF financial capability surveys, where focus groups were held in eight countries across three continents to establish a view of the attributes that distinguish someone who is financially capable from someone who is less financially capable. It also used the focus groups to identify the relative importance of knowledge, skills, attitudes, and behavior in distinguishing people with different levels of financial capability.

Given the exploratory nature of the work, the focus groups in this example were unstructured. Participants were asked first to describe someone who is very incapable with money and then someone who is very capable. The aim was to maintain about 20 minutes of discussion on each of these two topics to identify the range of competencies of importance and also the relative importance of knowledge, skills, attitudes, and behavior.

The respondents were all tea pluckers from the Muluzi Division in Bloomfied Estate at Lujeri (a population similar to the Direct Deposit project's survey sample). The participants were selected and asked to participate in the focus group by the division manager upon request from the research team. The first focus group had eight participants, five women and three men, while the second focus group had seven respondents and consisted of four men and three women. The focus groups were conducted outside the Muluzi division office.

Because researchers wanted to see whether a framework for measuring financial capability that had been developed in high-income countries had any relevance in a low-income setting, moderators ended the discussion by exploring the relevance of any areas that had not been mentioned in the earlier discussions.

In fact two areas or domains were consistently mentioned in the general discussion—day-to-day money management and planning for longer-term needs—while two other areas that had been identified in higher income countries—choosing suitable financial products and being well-informed—were mentioned much less frequently. When these latter areas were raised at the end of the focus group, they, not surprisingly, seemed to have much less relevance as core competencies for people with low incomes in low- or middle-income countries. It is interesting to note that there was a remarkably high degree of consensus in the groups across all eight countries.

What might focus group instruments look like to gather such information? The table captures the content of one such focus group instrument from an RTF pilot program in Nigeria. The six topics and questions were used to guide discussion in two focus groups, each lasting an hour.

(continued)

BOX 8.3 USING FOCUS GROUPS IN RTF PROGRAMS *(continued)*

TOPICS	QUESTIONS
Shopping Habits	▪ When do you usually buy groceries (payday)? ▪ When you buy on payday, do you buy food for the next two weeks or do you buy again in the shops during the week? Is it the same for vegetables as for maize? ▪ What other items do you buy on payday? Do you try and buy all of these also for the next two weeks? ▪ Do you shop at the payday market? Why or why not?
Loans	▪ If you do not have money when you need it what do you do? ▪ From whom do you generally borrow money when you need it? ▪ When during the fortnight do you generally take out loans (e.g., toward the end, in the beginning)? ▪ During the fortnight when are loans easier to get? ▪ When is the money from the loan paid back? ▪ How much time are you given to pay back? ▪ Is the loan pay back date tied to payday? ▪ What would happen if you didn't pay? ▪ What if people don't collect their wages on payday (e.g., when they are sick) but during the following week—how is the wage collected? ▪ How often do you buy store goods on credit? ▪ Do you pay interest for the goods bought on credit? ▪ Does the shopkeeper ever deny people credit? What were the circumstances?
Badiri	▪ How long do Badiri groups usually last? The whole season, less, or more? What is the reason? ▪ How are Badiri groups formed? ▪ How do you know the people in the group? ▪ How many people does a group generally consist of? ▪ Do people know what they want to buy with the savings before getting the savings? ▪ What was the purchase that you made from your last Badiri collection? ▪ Was this what you had planned to buy or not? ▪ What was the largest Badiri collection you have made in the last year? What did you purchase from this money? ▪ How frequently do people have to pay in? Is it tied to pay day? ▪ When do you pay in? Is it tied to payday? ▪ What happens if people don't want to contribute in a given week?
Savings	▪ Where is the money saved from payday, if any is left at all? (in addition to Badiri) (saving is not just explicit savings but any type of "storing")? ▪ Is there any money left over from the rainy season to spend now? If so, where was it kept? ▪ If you need for food/school fees, how do you acquire the money? Keep probing on this question: if you can't borrow from your friend, what else would you do? If you couldn't get extra wages, what else could you do? Etc.
Pres-sures to Share	▪ Do friends or relatives ask you for money? How often? ▪ Generally what reason do they give you for asking the money? ▪ For what reasons are you more likely to give money? ▪ When are they most likely to ask you for money? ▪ Share personal experiences when you were asked for money by your relatives. What was your response?
Security	▪ Do you worry about being robbed on payday?

gathering through focus groups can range from very structured to very unstructured. It may, for example, be more unstructured during exploratory research, enabling different themes to emerge spontaneously from the interactions among participants. Then again, more structured formats would be used when research on a topic has already been conducted and the goal is to elicit insights into specific questions arising from the researchers' previous findings. For instance, focus groups can provide context and detail to quantitative survey findings, allowing researchers to explore why certain results were obtained. Not surprisingly, focus groups in impact evaluations tend to be relatively structured because researchers already have a good idea of the issues they want to explore.

8.1.4 Desk review of documents and materials

Reviewing program documents can provide a rich source of information on a program's stated aims and objectives.

A key method used in most—if not all—program evaluations involves systematic document reviews: reviewing documents, records (including archival ones), and data associated with and relevant to the program under scrutiny. In program evaluations, there are different categories of potentially informative information, including:

- Official documents and materials describing the program's aims, structure, and so forth (including, perhaps, a program's website)
- Program materials not intended for public circulation (such as meeting minutes, internal progress reports, internal communications about the program, etc.)
- Data gathered in the course of implementing a program (for instance, demographic information about the program beneficiaries, results of specific activities, logs of program activities, etc.)
- Photographs and audio and video recordings
- Nonprogram data (such as financial transactions, school enrollment records, and so forth).

These materials often provide a very rich source of information on a program's stated aims and objectives, design, implementation, processes, stakeholders (funders, staff, clients/beneficiaries, etc.), inputs (including financial information), outputs, and results. They may also document changes that occurred in the program during its lifetime and that may be relevant to the evaluation. Figure 8.1 provides some examples of the types of documents and what they are useful for.

Program materials can also provide direct input to the assessment of the **quality** of a financial capability program. For example, the quality of program materials (inputs) can be reviewed and assessed by skilled peer reviewers, as can an audio recording of the program being delivered (outputs).

FIGURE 8.1 KINDS OF DOCUMENTS TO REVIEW

To describe the program's aims, objectives, and structure	■ Official and unofficial program materials (program descriptions, websites, meeting minutes, etc.)
To understand the social, economic, cultural, and political context in which the program operates	■ Official and unofficial program materials ■ Existing research (books and articles about the area) ■ Newspapers and other media
To understand program implementation	■ Official and unofficial program materials ■ Monitoring data reports
To gain insights into the program's beneficiaries	■ Monitoring data reports ■ Census and/or survey data reports ■ Media reports ■ Official and unofficial program materials

BOX 8.4 DESK REVIEW TO ASSESS FINANCIAL ADVICE

To illustrate, audio-recordings of a one-on-one financial capability guidance service being piloted in Great Britain were peer reviewed to assess the quality of the advice against a predetermined set of criteria, including:

- The accuracy of the information and guidance given
- Completeness of the information and guidance
- Whether the information and guidance given was appropriate to the needs and understanding of the user
- Various aspects of effectiveness, including whether the money guide went beyond the presented problem to identify the full extent of needs for help, whether they made certain that the user understood the information and guidance they were given, and whether they built up sufficient rapport to maximize the likelihood of the user acting on the guidance they had been given.

At the end of the checklist, automatic scores were calculated from the assessments under each of the broad headings and the cases allocated to one of three categories: Acceptable, Partially Unacceptable, or Unacceptable.

8.1.5 Audit or mystery shopping studies

Audits can provide information on how a program is delivered and whether it varies given beneficiaries' perspectives.

Another qualitative data collection method is audit or mystery shopping studies—an approach that can provide useful information on how the program is being delivered and whether delivery varies significantly depending on perceived characteristics of the beneficiary.

Audit studies have been used recently to study the behaviors of service providers in particular industries. This methodology consists of using trained auditors, or mystery shoppers, to pose as a user (e.g., client, customer, job applicant) to assess a financial capability program from the user's perspective and to study the response of the interlocutor (service provider, HR department, etc.).

Although often described as a qualitative data–collection technique, larger audits can provide quantitative data if they are able to include several hundred assessments. Usually, though, it is a technique that is used to provide in-depth information, and the number of audits is limited to between 30 and 50.

It is critically important for audit studies that the interaction remains "blind" and that auditors are truly "mystery shoppers," with their role as auditors concealed. Some studies use researchers or peer reviewers as the mystery shoppers, while others use real-life consumers or trained actors working with a script. The advantage of using researchers or peer reviewers is that they know what they are looking for in the evaluation and will thus ensure that all relevant information is collected, especially in situations in which scripts are difficult to develop and contingencies are hard to foresee. Using real-life consumers with the characteristics required for the evaluation

BOX 8.5 AUDIT STUDIES IN PRACTICE: AN EXAMPLE FROM AN RTF PILOT PROGRAM IN MEXICO

An RTF pilot program that evaluates the effectiveness of Mexico's credit disclosure reforms on consumer understanding and financial decision making uses audit studies. This project sends auditors, posing as customers, to financial institutions to seek information on savings and credit products. The researchers then compare the quality of information that auditors are given to information that is provided to regulators. An important benefit of audit studies is the ability to experimentally vary auditor traits, notably financial capability, to understand whether products and services are provided differently based on the perceived characteristics of the individual. This can have important implications for the success of targeting as well as potentially negative practices such as discrimination, bias, and fraud. For instance in the RTF pilot program in Mexico, auditors act as either inexperienced or experienced consumers.

has the advantage that they can, in effect "be themselves" and not have to act a part. They will, however, need some training (and/or a checklist) to minimize the risk of failing to collect all relevant information.

Interactions of auditors or mystery shoppers must be blind, with their roles concealed.

The experiences of shopper's interactions are generally documented in one of three ways:

- Audio or digital recording
- Detailed diaries of interactions
- A checklist provided by the research team.

The advantage of audio recording experiences is that there is a full record of interactions that can be transcribed. But they are really only suitable for a qualitative study. And they raise significant ethical issues, because the person being assessed must, by necessity, be kept unaware of the recording and for this reason it is generally ruled out.

Diaries are generally based on free text, so that they too are only really suitable for a qualitative evaluation. Compared with recordings, diaries suffer from the disadvantage that they record only the shopper's perceptions of what happened and information may not be verifiable. This approach is most appropriate where researchers are posing as mystery shoppers.

Mystery shoppers can record interactions, keep diaries, or complete checklists.

The third option is to develop a checklist for mystery shoppers to use in assessing the interaction, which should be completed immediately following the interaction. Often, the checklist is precoded. This makes subsequent analysis more straightforward and means that the checklist can be used for studies involving large numbers of audits. The key disadvantage is that the checklist needs very careful design to ensure it is comprehensive and does not influence the shoppers' behavior. This becomes more difficult if the assessment is complex or wide-ranging.

8.2 QUANTITATIVE DATA COLLECTION METHODS

Unlike qualitative data, quantitative data describe data that are either collected as numbers or that can be easily translated into numbers. In financial capability programs, statistical analyses can be conducted on quantitative data to summarize, compare, and generalize beliefs, attitudes, and behaviors. The key advantage of quantitative data is that they can be used to statistically test hypotheses and relationships. Where qualitative research can help to probe relationships between a financial capability program and observed outcomes, quantitative research can be used to describe, predict, and establish those relationships in the aggregate.

Quantitative data can be used to statistically test hypotheses and relationships.

Because of the quantitative nature of impact evaluations—with their emphasis on determining causative effects—quantitative data naturally go hand-in-hand with

impact evaluations. However, quantitative data can also be a very helpful part of monitoring systems and process evaluations. As in the example in chapter 4 on monitoring, collecting observational data on the ratio of functional computers to total computers or the number of individuals who complete a training course are examples of quantitative data a researcher might collect as part of monitoring a computer-based personal finance training session. As another example, in chapter 5 on process evaluations, we used the example that a school-based financial literacy program may want to conduct a short survey to understand how teachers feel about the course material, how students respond to the lessons, and what practical challenges exist in different contexts.

Quantitative data can be either primary or secondary data.

There are two broad categories of quantitative data: primary and secondary data. Primary data are collected firsthand by the researchers, directly from the subjects under study. Surveys, direct observation, experiments, and field experiments are all examples of primary data collection methods.

Secondary data are existing data that are collected by someone other than the evaluation team but that can be used by the evaluation team to help answer relevant questions. Administrative data or census data are examples of secondary data. In many cases, evaluators will link administrative data with survey data to provide a richer and more complete answer to questions about the effects of financial capability programs.

In the sections that follow we describe the most commonly used quantitative data collection methods used for monitoring and evaluation of financial capability programs: surveys and site visits (representing primary data collection) and administrative data (representing secondary data collection).

8.2.1 Surveys

Surveys are typically the best data collection method for determining information about large populations, with a known level of accuracy. In chapter 3, we discussed how to define a set of indicators that are program-relevant. But determining how to practically measure these indicators can be more challenging than it may appear.

Surveys are best for determining information about large populations, with a known level of accuracy.

In addition, while the primary objective of data collection is to enable evaluators to assess outcome indicators of interests, most primary data collection efforts typically cover significantly more than the outcomes alone. Collecting background information on respondents is important because it allows evaluators to better understand the study population, control for various sociodemographic characteristics, and look for heterogeneous effects.

It is also important to measure factors other than the program that contribute to financial decision making. As we discussed earlier in chapter 6, evaluators want to be

BOX 8.6 HOUSEHOLD SURVEY DESIGN IN DEVELOPING COUNTRIES

Typically, a household survey begins with a household roster to collect information on the household and is important in establishing who is a part of the household by listing all members. In addition to determining household membership, the household roster is used to collect information on the household members and those family members who do not reside in the household, if need be. This information usually pertains to age, gender, education, employment, and literacy.

When dealing with employment questions, it is necessary to ensure that the questions remain valid for workers engaged in a diverse range of economic activities, such as agricultural, nonagricultural, rural and urban self-employed, rural and urban employed, child laborers, unpaid workers such as those involved in family businesses, and those who receive nonlabor income. A second important component is the collection of information on consumption and expenditure, assets, and sources of income. Finally, other modules within the survey depend on research objectives. In the case of financial capability programs, this often involves questions related to risk and time preferences and social networks, as well as the responsibilities of financial decision making within the household. The choice of modules also depends on the length of the survey instrument and time constraints for responders; some household surveys can take several hours to administer or require several visits.

The design of household surveys is a complex undertaking, and the interested reader is referred to the survey instruments in appendix E (available in the online version of this report and accessible at www.finlitedu.org) for best-practice examples. Working with an experienced survey research firm can also help to ensure that the form, length, and content of household surveys conforms to accepted research practice standards.

able to tease out the effect of the program being evaluated from other things in the environment that could have also caused the effect. That means that survey instruments should be able to assess program outcome indicators of interest and capture information that is vital to understanding the indicators, such as demographic and household data on living standards, preferences, and behaviors, and the economic and social environment of the individuals or households of interest.

All of this argues that preparation for survey instrument design should begin by carefully reviewing both the indicators that evolve from the program's results framework and the experience of other data collection exercises that may have been undertaken in a similar setting. The framing, formatting, wording, and delivery of questions can all affect measurement. In general, as noted previously, questions should be as consistent as possible with best-practice surveys, subject to the condition that they adequately reflect the program components. To facilitate this, appendix E (available in the online version of this report and accessible at www.finlitedu.org) includes a number of survey instruments used by the RTF pilot programs and other best-practice examples, including the survey instrument developed by the RTF Measurement Project.

Survey instruments should be derived from indicators that come out of the program's results framework.

The advantage of using previously fielded, well-validated questions is clear. But the first priority for evaluation is to use questions that best capture the indicators for the program components in the setting at hand and to ensure that adaptation is appropriate, something we covered earlier in discussing interview questions. The diversity of financial environments—even within developing-country settings—ensures that while some aspects of the survey questionnaire may be commonly held, individual data collection efforts are still highly context-driven.

For instance, to measure planning for financial well-being in old age in a survey targeted toward workers in a **middle-income country**, researchers might want to ask about retirement savings. In that case, they might want to be specific in excluding government programs (e.g., "please do not include money that you expect to receive from government programs, such as Social Security"). In contrast, to measure planning for financial well-being in old age in a survey in a **low-income country**, it may make more sense and be more appropriate to be more general and ask the respondent how he plans to be financially supported in old age rather than pose questions about retirement savings.

Relevant response options for both countries may still include savings through formal financial institutions, such as retirement or bank accounts, and should also take into account (possibly by name) any widely used existing government or nonprofit schemes. In certain cultures, savings clubs may be popular and have specific names. Options should also cover informal arrangements, such as depending on children or younger relatives for support.

Care must be taken when adapting a survey instrument from other evaluations, even those that cover the same types of outcomes in similar settings.

The bottom line is that care must be taken when adapting an instrument used by other evaluations, even those that cover the same types of outcomes and interventions in nominally similar settings. As a final note, questions that are translated from English or another language into local languages for respondents should be back-translated independently, and the back-translation should be checked to ensure that the translation captures the intent of the original question.

No survey—regardless of how relevant—is going to be "turnkey." That means there will likely be a need to develop new questions. In thinking about question development, there are two types of questions: open-ended ones and closed-ended ones. An open-ended question allows respondents to provide a verbal or written response of their choice, while closed-ended questions require the respondent to choose from prespecified answers. The former allow the respondent more freedom in answering a question, allowing for a richer set of responses, while the latter, though restrictive, are generally easier to code, clean, and analyze.

It also makes sense to take some additional time and scrutinize question performance. As was true with interview questions, we use the process of cognitive

testing, where in-depth interviews with interviewees are used to ask respondents to describe their understanding of the question and thought processes in coming up with answers. It is also very instructive to ask interviewees to restate the question in their own words. The focus of such cognitive testing is to understand if questions correctly capture the researchers' underlying intent by assessing whether respondents are likely to understand questions correctly, whether the responses provided are appropriate and inclusive, and whether there are potential common errors. We note that cognitive testing is inherently qualitative (as described earlier) and thus does not need to involve more than a small number of purposively selected subjects.

How questions are ordered or positioned in a survey questionnaire can affect participants' responses.

When putting a survey instrument together, questions used should be identical across all respondents (especially treatment and comparison groups, before and after an intervention) to ensure that biases are not inadvertently introduced. The order of questions should be explicitly considered as well, because this can also affect response. One reason is **saliency**: For instance, asking about perceptions of financial status before or after a battery of questions about income and employment may result in different responses because individuals may better recall certain sources of income in the latter case.

Another reason is **inadvertent framing**: Residual positive or negative emotions from a sensitive series of previous questions can affect response, or even result in refusal to further participate. Many household surveys put financial decision making at the end of the survey modules because of its potential sensitivity; however, this should be balanced against considerations of length and fatigue.

Table 8.2 provides a checklist of things to consider in designing a survey instrument.

TABLE 8.2 CHECKLIST FOR SURVEY INSTRUMENT DESIGN

Make the questions short and in simple, everyday language, using terms that are specific and leave little room for interpretation	X
Make the questions relevant to the culture, environment, and experiences of the population of interest	X
Write the question in full so interviewers need only read them, rather than introducing their own interpretation	X
Given a complex survey, have closed-ended questions with answers precoded as far as possible and with coding uniform across all questions	X
Make sure to include clear and explicit skip patterns (i.e., guidance on when to skip from one question to another when a respondent answers in a certain way)	X

8.2.2 Site visits and observation

Sites visits and nonparticipant observation can be either qualitative or quantitative or both, but they are, under all circumstances, very important elements of program

evaluations. During a site visit and/or observation, researchers visit a program site and objectively and carefully observe a program's operations and its interactions with participants as they are taking place.

The insights gleaned from this kind of data collection can provide a useful and informative complement to the insights obtained from surveys, focus groups, or interviews, which report primarily on participants' perceptions, subjective experience, behaviors, attitudes, and beliefs. And they also provide important information that is not necessarily explicit in official descriptions of the program or that may not come up in the context of interviews, focus groups, or surveys. This includes information about context, processes, and interactions.

Site visits and observation do not typically require a researcher to interact with subjects or to play an active role in what is taking place (unlike **participant observation**, a mainstay of anthropological inquiry). Rather, the researcher is required to observe a phenomenon and note its characteristics descriptively.

Such a description can include a number of elements, including the setting where the program is being implemented, the people receiving the intervention, the interactions between those individuals, the actual activities that occurred (which can be compared with what was intended to be delivered), and factors that might have enhanced or detracted from the program. In terms of the latter, the skill or lack of skill of whoever is delivering the program activities can be witnessed firsthand. And, of course, such descriptions can also be used as one basis for peer-reviewing program quality.

Regardless of the type of information collected during a site visit and observation, it is important to collect it systematically, using either a template or topic list that has been compiled in advance. If observation techniques are used, it is very important that the people being observed are put at ease and feel comfortable with the observation; otherwise, what they do will be influenced by their unease, and the observations will be invalid.

Site visits allow program evaluators to get inside the "black box" of program activities.

Site visits and observation can help program evaluators get inside the "black box" of program activities and see for themselves what is actually going on. How are the program activities being delivered and how are the participants receiving them? While site visits may not in themselves be sufficient sources of information about a program, triangulating these descriptions with other data (such as from surveys, interviews, focus groups, and document reviews) can provide rich and nuanced insights for an evaluation.

BOX 8.7 USING ADMINISTRATIVE DATA IN PRACTICE: RTF PROJECTS

In chapter 4 (Monitoring), we discussed the RTF pilot program in Malawi that aims to encourage formal savings by agricultural workers by directly depositing wages into bank accounts. This program uses administrative data from bank records in addition to surveys to measure changes in consumption patterns as well as formal and informal savings and borrowing behaviors among workers from poor households. These data are used to evaluate the impact of paying wages using a direct-deposit system, rather than cash, on saving behaviors and levels.

Another RTF pilot program based in Brazil works with the Brazilian stock market (BM&FBOVESPA) to test a range of interventions aimed at understanding the psychology behind investing behavior and biases and how to better improve financial decisions through targeted education. Administrative data from BM&FBOVESPA are used to test how financial capability might influence participation in new markets and investor biases in markets. Historic data on stock market dropouts will be used to identify common investor mistakes and biases that might cause investors to limit investments.

8.2.3 Using existing administrative data

In the previous two sections, we have talked about collecting data—either qualitative or quantitative—to support monitoring and evaluation. In all cases, the data collected are specific to the financial capability program being implemented. We refer to this activity as primary data collection.

But the reality is that data are being collected all the time, independent of the data being collected to evaluate a financial capability program, and those data can be helpful in supplementing the primary data collection activities. It is common to merge or link administrative data sets that already exist with survey data that are being collected as part of the evaluation process. For example, it may be possible to link self-report survey responses to administrative data to verify the self-reports.

Box 8.8 summarizes the types of existing data that are typically most useful in financial capability monitoring and evaluation.

BOX 8.8 ADMINISTRATIVE DATA FOR EVALUATING FINANCIAL CAPABILITY PROGRAMS

Household survey data. National household surveys are periodically conducted in many developing countries. These include multitopic surveys, such as the Living Standards Measurement Survey and the Demographic and Health Survey, which can cover a wide range of information on housing characteristics, household consumption and wealth, individual employment, education, and health indicators. Other surveys, such as labor force surveys, are more restricted in scope and sometimes cover only urban areas.

Where to look:

- Statistical institutes in the respective country
- International Household Survey Network (www.ihsn.org)
- Demographic and Health Surveys (http://www.measuredhs.com/)
- Living Standards Measurement Surveys (http://iresearch.worldbank.org/lsms/lsmssurveyFinder.htm)

Census data. Most countries conduct a population and housing census every 10 years, and many conduct additional surveys. The advantage of census data is that they cover the entire population, so there are data for virtually every potential treatment and comparison observation. The drawback of census data is that it is infrequent and typically contains only a limited number of indicators, limiting their value for an impact evaluation.

Where to look: International Household Survey Network (www.ihsn.org).

Facility survey data. Facility surveys collect data at the level of service provision, such as at a school or vocational training center. National ministries, state entities, or even local authorities may compile the information. In many cases, facility-level surveys will provide control variables (such as teacher–student ratio), while others may capture outcomes of interest, such as attendance rates.

Where to look: Relevant national ministries and local representatives.

Specialized survey data. A specialized survey is one that is collected for a specific purpose, often for research on a particular topic. Many take modules from the existing national household survey and add questions on topics of interest. Coverage of specialized surveys can be quite limited, sometimes resulting in little or no overlap with program areas. Nevertheless, if the evaluation team can find existing data from a specialized survey on a topic related to the evaluation, these data sets can provide a rich collection of relevant indicators.

Where to look: Local officials, donors, and NGOs in the area of interest.

Source: Extract from Hempel and Fiala (2010).

8.3 COMPARING THE USES OF DIFFERENT DATA COLLECTION METHODS

Above we discussed some of the key types of qualitative and quantitative data collection methods that are typically used in all evaluations, including those for financial capability programs. Table 8.3 summarizes the methods presented in terms of what they are and provides some notes about their uses and limitations in conducting evaluations.

KEY POINTS

Monitoring and evaluation efforts depend on the data collected to implement them. Such data can either be qualitative (such as focus groups with participants) or quantitative (such as surveys) or, more likely, a combination of both types. What types make sense—and in what combination—will obviously depend on the program objectives and activities. And doing it well and getting the most value out of the data collection effort will also depend on **how** those data are collected—a topic we cover in more detail in the next chapter, where we provide some practical tips on how to collect the various types of data.

TABLE 8.3 COMPARING KEY QUALITATIVE AND QUANTITATIVE DATA COLLECTION METHODS

METHOD TYPE	DESCRIPTION	NOTES ABOUT USES AND LIMITATIONS IN EVALUATIONS
Qualitative data		
In-depth interview	Uses topic guides that cover range of topics to allow interviewer to explore in depth expectations, experiences, views, and feelings of interviewee	▪ Is useful for discussing in detail personal information or sensitive issues that people may be hesitant to discuss in a group setting ▪ Lacks protocol, which limits comparability of responses across interviewees
Semi-structured and cognitive interview	Uses interview protocols providing prompts to allow interviewer probes about specific questions and issues	▪ Is useful because use of protocol increases comparability of responses and enables clearer insights into particular issues of interest
Focus group	Uses a moderated discussion of a particular topic or topics with a small group of people (usually between 6 and 10)	▪ Can be: – Cost-effective – Used to obtain information from transient populations – Used to spot areas of shared or diverse experience – Useful when informants are too intimidated or otherwise unwilling to have one-on-one interactions ▪ May raise new questions or point to aspects of the program being evaluated ▪ May be unsuitable to explore certain socially sensitive or highly personal issues
Desk review of documents and materials	Involves systematic reviews of documents, records (including archival ones), and data associated with and relevant to the program under scrutiny	▪ Provides rich source of information on program's stated aims and objectives, design, implementation, processes, stakeholders, inputs, outputs, and results ▪ May document changes in program during its lifetime relevant to evaluation
Audit study or mystery shopping	Uses trained auditors, or mystery shoppers, to pose as a user to assess a program from user's perspective and to study the response of the interlocutor	▪ Can provide useful information on how program is being delivered and whether delivery varies significantly depending on perceived characteristics of the beneficiary ▪ Must be carefully designed so interactions remain "blind" and so auditor role is concealed
Quantitative data		
Survey	Uses structured sequence of questions to collect background information on program participants and data on outcome indicators of interests	▪ Is best method for determining information about large populations, with a known level of accuracy ▪ Can assess program outcome indicators of interest and capture background information vital to understanding indicators ▪ Must be designed carefully to avoid response biases and respondent burden ▪ Has limits because it relies on self-reported data
Site or observational visits	Involves visiting program site and objectively and carefully observing program's operations and its interactions with participants as such interactions occur	▪ Can help program evaluators get inside "black box" of program activities and see what is actually going on ▪ Can provide useful complement to insights obtained from surveys, focus groups, or interviews ▪ Can provide important information not necessarily explicit in official program descriptions ▪ Must collect information systematically, using either a template or topic list compiled in advance
Existing administrative data	Uses secondary data collected for nonprogram purposes (e.g., Census data)	▪ Can supplement primary data collection activities ▪ Can be merged or linked to survey data and used to verify such self-reported data

FURTHER READING

General

Hempel, K., and N. Fiala. 2010. *Measuring Success of Youth Livelihood Interventions: A Practical Guide to Monitoring and Intervention*, Washington, DC: Global Partnership for Youth Employment.

Iarossi, G. 2006. *The Power of Survey Design: A User's Guide for Managing Surveys, Interpreting Results, and Influencing Respondents*, Washington, DC: World Bank.

Leeuw, E. D. De, J. J. Hox, and D. A. Dillman. 2008. *International Handbook of Survey Methodology*, New York: Lawrence Erlbaum Associates.

Morgan, D. L. 1997. *Focus Groups as Qualitative Research*, Thousand Oaks, CA: Sage Publications.

Technical

Campbell, D. T., J. C. Stanley, and N. L. Gage. 1963. *Experimental and Quasi-Experimental Designs for Research*, Boston: Houghton Mifflin.

Johnson, B., and Turner, L. A. 2003. "Data Collection Strategies in Mixed Methods Research," *Handbook of Mixed Methods in Social and Behavioral Research*, pp. 297–319.

Maynard-Tucker, G. 2000. "Conducting Focus Groups in Developing Countries: Skill Training for Local Bilingual Facilitators," *Qualitative Health Research* 10 (3): 396–410.

Sim, J. 2001. "Collecting and Analysing Qualitative Data: Issues Raised by the Focus Group," *Journal of Advanced Nursing* 28 (2): 345–52.

Shadish, W. R., Cook, T. D., and Campbell, D. T. 2002. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*, Houghton Mifflin College Div.

CHAPTER 9

The process of collecting data: practical guidance

In the previous chapter, we discussed the types of data collection methods—both qualitative and quantitative—that are key to enabling evaluators to assess financial capability programs.

While the preceding chapter described the types of data collection methods, it did not describe the **processes** of actually collecting the data. Those processes are central to getting the most out of the various data collection methods and therefore critical in ensuring the quality of any evaluation. How does one actually conduct a focus group? How does one select an appropriate sample population for a survey to ensure that the analysis of the resulting data will be both valid and useful?

The process of collecting data is key to getting the most out of data collection.

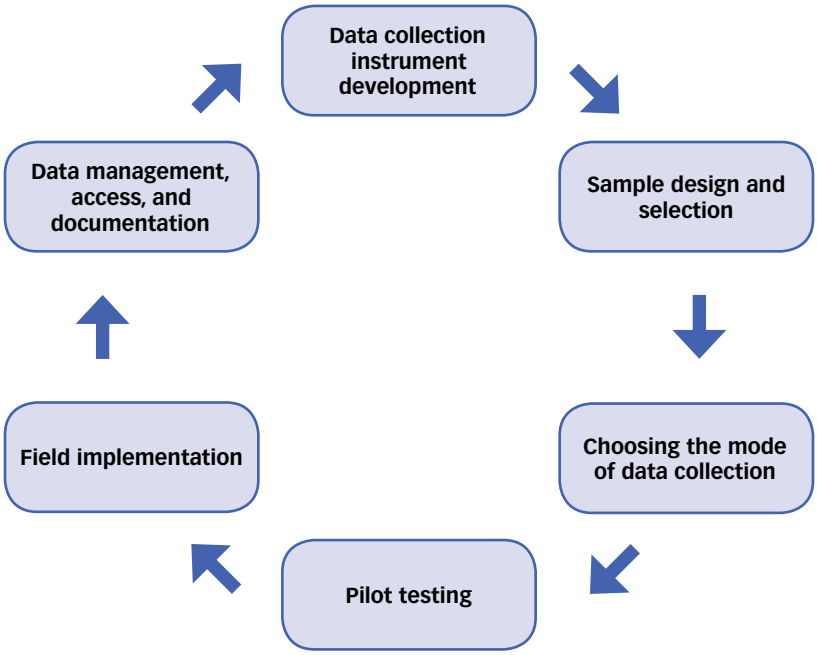
In this chapter, we answer questions like these by discussing some practical techniques to collecting data using the qualitative and quantitative data collection methods described in chapter 8.

9.1 THINKING ABOUT THE PROCESS OF DATA COLLECTION

While some evaluations may be conducted based on secondary data alone, in most cases, evaluations of financial capability programs will typically require collecting primary data from participants. Figure 9.1 shows the typical cycle or process of data collection in an evaluation, starting from instrument development, sampling, choosing modes of collection and pilot testing, fielding the data collection effort, managing the data that is collected, and documenting the data. While not all these steps are equally applicable in both qualitative and quantitative research, some are, such as sampling.

We discussed instrument design in chapter 8. We organize the remainder of this chapter around the remaining steps in the cycle, presenting some concrete guidance in performing those steps to ensure that the data collection effort is relevant, efficient, and of high quality.

FIGURE 9.1 THE PROCESS OF DATA COLLECTION

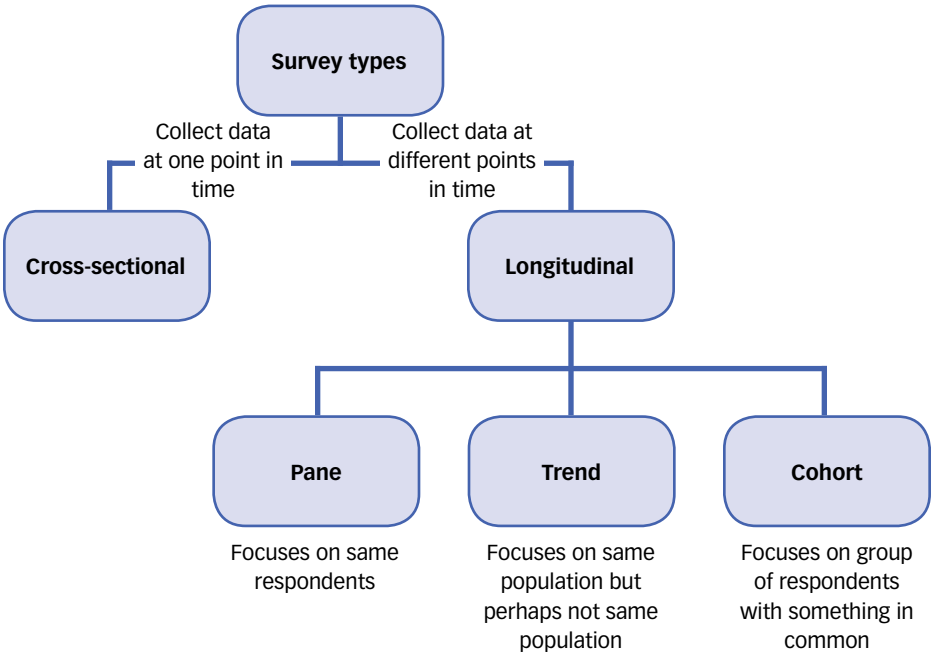


A key step in survey instrument design is to decide whether the survey will be a cross-sectional survey or a longitudinal survey, which differ in terms of how information is collected. In a cross-sectional survey, information is collected from respondents at a single point in time, whereas information is collected over a period of time in a longitudinal study.

There are three broad types of longitudinal studies: **trend** studies, **cohort** studies, and **panel** studies. In a trend study, respondents from the same population of interest are surveyed repeatedly, at different points in time. While the population is the same, the respondents in the survey may vary each time. A cohort study also surveys the same population at different points in time, but it is slightly different because it focuses on a cohort, a group of people with something in common. A panel study collects information at various points in time using the same respondents. Figure 9.2 summarizes this distinction graphically.

For example, an annual financial literacy survey of recent high school graduates would be a trend study, while a cohort study of financial literacy may sample from children born in 2005 and survey those children at 1st, 5th, and 9th grades and upon high school graduation. And a panel study may follow a particular class of 1st graders, and survey those same students at 5th, 9th, and 12th grades.

FIGURE 9.2 SURVEY TYPES



9.2 SAMPLING DESIGN AND SELECTION

Not surprisingly, budgetary and time constraints may not permit evaluators to collect detailed information on **everyone** in a financial capability program and the comparison population that is not in the program. The solution to this issue is sampling—the process of selecting units, such as individuals or organizations, from a larger population. Drawing a good sample means that the chosen sample accurately represents the entire population of interest; this, in turn, means that the conclusions about outcomes that evaluators make from analyzing the survey data can be considered valid. Sampling design and selection is one step in the data collection process that applies to both quantitative data collection (e.g., surveys) and to qualitative data collection (e.g., in-depth interviews and focus groups), although sampling is a much more formal part of quantitative data collection. We discuss some of the practical issues for both below.

Sampling is a more formal part of quantitative data collection.

9.2.1 Sampling for quantitative research

Not surprisingly, budgetary and time constraints may not permit evaluators to collect detailed information on **everyone** in a financial capability program and the comparison population that is not in the program. The solution to this issue is sampling—the process of selecting units, such as individuals or organizations, from a larger population.

Choosing the sample requires balancing cost, time, survey complexity, and accuracy.

BOX 9.1 CHALLENGES OF OBTAINING A SAMPLE FRAME

In some cases, the population of interest is straightforward and easy to list—such as the universe of all existing clients in a large bank or the universe of students in a school cohort. However, in practice, obtaining a complete and up-to-date sample frame for a population of interest is not always so straightforward. For example, in the case of households, census data may be outdated or not available. In these cases, information about the sample frame itself may need to be collected. For instance, in a prospective evaluation, there may be no census of individuals in a particular catchment area, which may mean that program staff must physically count all the individuals in that area—something that is called an enumeration or listing process. Here, having relationships with stakeholders, including local researchers, or working with an established survey research firm can be extremely helpful.

Choosing the sample is one of the most challenging aspects of data collection—one that almost always requires balancing considerations of cost, time, survey complexity, and accuracy. While the guidelines above apply generally to most settings, in practice, the process of sampling is highly context-specific and should be approached carefully in every survey by an experienced technical expert.

In thinking through the possible sample designs, evaluators have a number of choices:

- **Random or probability samples**: The best way to approach sampling is to draw individuals **randomly** from the population of interest. In this design, the sampling expert would first define the population of interest and obtain a "sampling frame"—a complete list of observational units from which one can draw a sample. A large-enough sample of random draws from the sampling frame will produce a survey sample that is likely to closely represent the characteristics of the larger population.

  For example, for a representative national household survey, a list of all households based on the most recent population census can serve as the sampling frame; for a representative sample of bank customers, this may be a list of all registered accountholders.

- **Systematic sampling**: Here, sampling involves the selection of elements from an ordered sampling frame.

  For example, microcredit clients may be ordered alphabetically and every third person selected for the sample.

Once a basic sample design is chosen, that design can be modified to provide desirable characteristics that aid data analysis or balance data quality with data collection costs, using two common techniques:

A chosen sample design can be modified through different techniques.

- **Stratification:** By stratification, we mean dividing the population into predefined **strata** or certain subpopulations or categories of interest and then randomly sampling a fixed number of units from within those subgroups. This allows data users to "guarantee" representation of subpopulations that might otherwise be ignored in a simple random sample because their numbers are too small to be accurately represented if members of the population are selected at random; it also increases efficiency by allowing the survey to focus on populations of interest more directly.

  For instance, a sample of bank customers could be stratified by gender and region to ensure the representation of men and women and urban and rural customers.

- **Cluster sampling:** In a cluster sampling method, groups such as villages or neighborhoods, are first chosen through a random sample, and then observational units are randomly sampled from within those clusters. This is a form of **multistage sampling**.

  For instance, instead of randomly drawing bank customers from a district roster, you may choose to randomly draw a set of villages from the list of villages in a district and then randomly draw bank customers from within each selected village.

  Because costs are almost always a limiting factor when implementing surveys, the benefit of cluster sampling is pretty obvious: Observational units within the clusters and the clusters themselves may be less costly to access; it is usually cheaper to travel to a handful of selected villages and locate a group of individuals in that village than to travel to a large number of villages and locate one or two individuals in each. The drawback is also pretty obvious: many people in a given village may be quite similar to one another and their outcomes may be highly correlated, reducing the usefulness of the data from the sample as the evaluator gains only a limited amount of new information. It is generally best to have as more clusters that are smaller in size than the reverse.

The question of sample size in quantitative analysis is very formalized. Evaluators need an explicit **power calculation** to derive estimates of the sample size for each design needed to convincingly distinguish between a lack of actual impact and an inability to statistically detect meaningful differences. In general, a larger sample is needed if (a) the expected effects on the outcome are small, (b) the number of

subgroups studied is large, (c) the level of statistical confidence required is high, and (d) the chance of failing to detect an actual effect is low. Since individual power calculations are based on a single outcome, they should be conducted on the primary outcome indicators for the program. Various statistical packages allow evaluators to make such power calculations. Your sampling expert's role is to articulate scenarios that are determined by the demands of the formal statistical models and assumptions about the program background to select a final survey design and sample size—a task that balances logistical considerations against the precision needed to capture reasonable program effects. Sampling is discussed in further detail in appendix B.

Finally, for quantitative analysis, because data collected for impact evaluations are intended to make inferences about the larger population of interest, external validity adjustments may need to be made through **sampling weights** so the sample can be rebalanced to look like the community of interest. If certain segments of the population were more likely to be selected for the sample (by design from stratification, or other factors such as selective nonresponse), the sampling specialist should help determine sampling weights that reflect the probability of being sampled, so that evaluators can produce estimates that represent the full population of interest.

The advantages and disadvantages of each sampling method are summarized in table 9.1.

TABLE 9.1 BENEFITS AND LIMITATIONS OF SAMPLING METHODS

METHOD	ADVANTAGE	DISADVANTAGE
Simple random sample	▪ Simple procedures and analysis	▪ Logistical difficulties, especially if population is far-flung ▪ Requires sampling frame
Stratified sample	▪ More likely to be representative, especially of subgroups ▪ Smaller total sample size required	▪ Need to know in advance relative sizes of strata in population ▪ Can be logistically difficult ▪ Requires weights in analysis
Cluster sample	▪ Logistical advantages	▪ Larger sample needed ▪ Requires adjustment to standard errors in analysis
Systematic sample	▪ Simple procedure ▪ Does not necessarily need to have sample frame	▪ Difficult if population is dispersed

Source: Levinsohn et al. (2006).

9.2.2 Sampling for qualitative research

When it comes to doing qualitative research, sampling is still important, but there are no widely accepted rules about adequate sample sizes. Because the aim of qualitative research is to provide a depth of contextual understanding, rather than statistics that can be generalized to the population, sample sizes and design can vary depending on the specific aims, circumstances, and resources of a particular study.

In most cases, sampling is **purposive** or **purposeful,** or based on participants' potential for providing the kind of information sought in the research. There are many types of purposeful sampling used by qualitative researchers, including random sampling. For instance, in the evaluation of a national program, a television soap opera with a financial capability storyline, such sampling could consist of the purposive selection of a particular set of communities, followed by the random selection of individuals within those communities. This type of sampling aims to increase the credibility of study design and findings, not for representativeness.

Another useful strategy is **maximum variation sampling**. In this strategy, cases exhibiting significant variation along one or more variables of interest are selected so multiple perspectives and experiences are represented in the research. This is perhaps the most common purposeful sampling approach. Qualitative research is often undertaken with the people responsible for planning and delivery of a financial education program—the implementers. Here, if the numbers are small, you may include everyone involved, either by interviewing them in-depth individually or by soliciting them as participants in focus groups. If the numbers are too large for this approach, then you will almost certainly want to select a cross-section of the people involved in different ways in the program. Another example of where it might be used is a situation wherein a quantitative impact survey has identified very different levels of impact.

Other types of common sampling strategies include **outlier sampling**, or the selection of cases based on their being unusual or outside the norm in a particular respect. For instance, in a financial capability evaluation, one may seek out the sites (or participants) where the program was particularly successful or unsuccessful in achieving a result, both of which may hold especially important lessons for the program. Finally, in some settings it can be desirable to recruit via referrals from a small sample of initially identified individuals to others in their social networks or connections. This type of sampling is called **snowball sampling**.

Instead of formal stratification as previously discussed, in qualitative analysis, a decision will need to be made at the outset on the key characteristics that the sample of people selected will need to possess and the numbers of people required with these different characteristics. This is known as "setting a quota." For example, in

BOX 9.2 CONVENIENCE SAMPLING

For both qualitative and quantitative research, the most common but least desirable approach to sampling (because it is less rigorous) is a "convenience sample"—also known as a "grab sample"—in which the sample is simply taken from an easy-to-access population. This could involve interviewing people waiting at a bank to assess knowledge of and demand for financial products. In many settings, this approach, while inexpensive and sometimes informative, can produce unreliable data. For instance, people waiting at a bank in a busy urban location may not be representative of the entire population of bank customers.

Unless cost and time are the only factors driving sample selection, samples that are chosen based on convenience alone should be avoided. Even though cost and time constraints are important considerations, it is critical to carefully construct a sampling strategy that enables the most information-rich cases to be studied.

recruiting for the interviews used in the development of the Russia Financial Literacy and Education Trust Fund (RTF) baseline survey of financial capability, it was decided to use a maximum variation sampling approach, with broad quotas set for income (roughly two-thirds low income; one-third middle/high income); sex (roughly equal numbers of men and women); work status (roughly equal numbers of people who worked informally, who were in formal employment, or did not work at all); and age (with roughly equal numbers in each of three age groups: under age 30, between ages 30 and 50, and over age 50). A common mistake is to set firm quotas for too many characteristics; doing so may make the process costly to find the last few people, because the combination of characteristics might be very uncommon. To overcome this difficulty, one typically either sets broad quotas, as above, or one sets exact quotas for a small number of characteristics (e.g., just income and age) and softer quotas for the other characteristics of interest (e.g., sex and work status), which are often expressed as "not more than x, but not fewer than y" people.

There are several methods that can be used to select the sample for a qualitative analysis. First, it can be from the participants in a quantitative survey of users if one is being included in the evaluation—as in the evaluation of the financial education program on general money management for people on low incomes (topic guides were included in chapter 8). Or it can be a sample that is selected directly from the population of interest, which is the approach used in the Russia Financial Literacy and Education Trust Fund cognitive interviews for the baseline survey of financial capability (also described in chapter 8). Wherever feasible, the former approach is preferred; wherever there is a previous quantitative survey to sample from, that will provide a wealth of information to inform the selection. If a sample has to be

selected directly from the population, this will usually require a short screening questionnaire in order to identify the people of interest.

Compared with quantitative evaluations, sample sizes in qualitative evaluations tend to be smaller in order to allow evaluators to conduct a detailed analysis of their content. Generally, the number of in-depth interviews of qualitative mystery shoppers or focus group participants will depend on the diversity of groups that is the focus of the evaluation.

For example, an evaluation of a financial education program delivered through workshops to young women with only basic levels of education will require smaller qualitative samples than one where financial education is delivered in three different ways to separate subgroups of such young women or one delivered through workshops to three very different groups of people: young women with only basic levels of education; young women who have completed high school education; and young women who have completed college, university, or other tertiary level education.

Where workshops of young women with only basic levels of education are evaluated, 30 in-depth interviews or mystery shoppers and four or five focus groups should provide the level of information required for a qualitative element of an evaluation, especially if it supplements a quantitative approach. In the other two instances, larger samples will be required, with 15–20 people in each of the subgroups for in-depth interviews or mystery shopping and three focus groups per subgroup. For this reason, it is important to keep the number of subgroups to a minimum.

With focus groups, there is also the additional consideration of determining the optimal number of participants. Most qualitative research experts agree on between 6 and 10 people, with eight perhaps being the ideal number. This will ensure that you have sufficient people for a full discussion but not too many people that the group becomes too difficult to manage and ends up being dominated by a small number of participants.

A further consideration in focus groups is the degree of diversity to include in a single group. Too little diversity and the discussion may be limited, but too much diversity could lead to the risk of some participants dominating the discussion. This could be the case in the example given above of an evaluation of financial education workshops to three groups of young women with very different levels of education. Separate groups for women with each of the three levels of education are likely to be more appropriate. Other instances where it can be advisable to hold separate groups include people with different income levels, people of different ages, and (in some societies) men and women. There are no hard-and-fast rules on this; it depends on the context of the evaluation, including the subject matter and the community in which the fieldwork is being done.

Like sampling for quantitative research, sampling for qualitative research is always a balancing act—in this case, between ensuring that you have captured sufficient diversity and keeping the amount of information collected to a manageable level for full qualitative analysis. The ideal, obviously, is to achieve a sample size where no new insights are coming to light and further fieldwork could not be justified. It is important to provide a solid rationale for the choice of sample size, paying particular attention to the goal of data saturation. Unlike in quantitative studies, which aim to generalize from a sample to the population, samples in qualitative research are intended to maximize information; the sample size should strive to reach the point at which the information obtained becomes redundant.

9.3 CHOOSING THE MODE OF DATA COLLECTION

Choosing the mode of data collection is another decision that often embodies trade-offs and requires careful consideration of specific local conditions. Data collection modes can affect the quality of data, but can vary significantly in cost and feasibility. Common modes are self-reporting through mail-in paper forms, telephone interviews, face-to-face interviews, and online surveys. Access to respondents through mail, telephone, face-to-face interviews, and the Internet may also vary significantly in different settings and lead to sample selection bias.

With surveys, mail-in forms are often the cheapest to generate and allow individuals privacy in providing sensitive data, but they typically result in low response rates. The success of postal surveys can vary given the state of the postal service. Telephone interviews or face-to-face interviews allow the interviewer to interact and motivate response, but they can also affect the information, because some people may give answers that they think are socially desirable (the "response bias"). Internet surveys are often relatively inexpensive, but the identity of the true respondent can be hard to verify, and, unlike with a face-to-face interview, getting a sense of how engaged respondents are is problematic. As with postal surveys, the low level of Internet penetration may make it unlikely that Internet household surveys at this time will generate useful information, except for situations such as businesses where Internet penetration is high. The use of paper-based forms can result in recording and transcription errors, but such forms may be the most feasible approach in low- and middle-income countries. In general, electronic data entry and transmission encourages accuracy. A recent innovation is computer-assisted personal interviewing (CAPI), in which enumerators travel with lightweight computer laptops or personal digital assistants (PDAs). In field tests, CAPI has been shown to result in much higher-quality data because of automatic error checking and fewer transcription errors. The optimal choice often depends on the budget and purpose for which the data are being collected.

When it comes to interviews and focus groups, an in-depth interview will normally last up to an hour if it is conducted face-to-face, although it could be longer if it is being held with someone involved with the planning or delivery of the service, as opposed to a participant. It is possible to conduct an in-depth interview by telephone—particularly with people involved in program planning or delivery—but the length will need to be shorter; as a rule of thumb, it should be about half an hour or so. Telephone interviews are often used for follow-up interviews after an initial one conducted face-to-face.

Whatever mode is chosen, it is important to consider maintaining the mode of data collection across respondents and waves, unless there is a compelling logistical constraint. For instance, budget and time constraints may lead evaluators to use a phone survey in urban areas, a face-to-face survey in rural areas, or a detailed face-to-face survey at baseline and a telephone follow-up down the road. In such cases where the data mode is not consistent, potential mode effects (discussed later) should be carefully considered in the analysis.

9.4 PILOT TESTING

Before going into the field in force with a data collection effort, it makes sense to pilot-test the data collection instruments and procedures first. Earlier we described **cognitive testing**, which generally involves the qualitative testing of instruments and protocols before finalizing a data collection instrument; **pilot testing** is different and refers to testing the instrument in the field among the actual program's target population. Note that some people also use the term **pretesting** or **field pretesting** to mean piloting. It is easy to underestimate the importance of pilot-testing instruments and procedures in the field before deployment. Time and budget constraints sometimes make testing sound like a "waste of precious time." But in almost all cases, even minimal testing can help evaluators refine the instruments they will use in their research and prevent problems later on, such as respondents not understanding questions or protocols failing to draw the information evaluators are after.

To prevent these issues, pilot-testing usually consists of actually fielding the survey instruments and interview and focus group protocols with a sample that resembles the target respondents as closely as possible to check for issues covered qualitatively during cognitive testing (such as clarity in the wording of specific questions) and, more generally, issues related to the overall administration of the instruments. Such issues include the flow of the instrument or protocols; the time it takes to complete the survey or interview, and whether the questions are adequate to obtain the necessary information.

For qualitative pilot testing, if the interviewers are not members of the research team, then the research team will want to hold a detailed briefing for the interviewers to provide them with sufficient background information such as the interview goals, topics to probe, and when to gather more details on points that are not in the topic guide but are important to capture for the evaluation.

Once you have piloted the evaluation tools, you can make the necessary adjustments. Typically, the responses obtained during piloting in quantitative evaluations are not used in the final evaluation. But it is not uncommon to use data from pilot interviews or focus groups in the final evaluation, as long as the content and types of participants do not deviate from what will be used for the study.

9.5 FIELD IMPLEMENTATION

Evaluations typically require managing a large team, including enumerators, supervisors, programmers, data entry operators, support staff, and others. Prior to launching any field operations—quantitative or qualitative data collection—preparing and finalizing an overall **workplan** ahead of time with all participants is a key part of ensuring smooth coordination.

In both qualitative and quantitative data collection, the formal part of any interaction begins with obtaining the respondent's informed consent to participate in the interview. It is important that the respondent understands before the survey or interview starts why he is being interviewed, how he was selected, how long it will take, who has commissioned and paid for the research, and how the results will be used. We discuss this in greater depth in chapter 13 (Ethics), where we talk about issues of getting consent.

In qualitative research, the nature of the interaction with the respondent is critically important. Because in-depth interviews call for a high level of skill it is important that those facilitating them have substantial interviewing experience, either in a research setting or some other context involving nondirective interviews (i.e., interviews that are allowed to follow the course the interviewer may set). It is very important that the interviewers do not influence what the respondent says and, above all, that they allow and encourage the respondent to speak at length on the topics to be covered. Interviewers should have well-developed listening skills and be familiar with techniques to probe replies and encourage the respondent to elaborate, such as using neutral prompts like "Why do you say that?" and "Can you tell me more about that?" A good in-depth interviewer will allow respondents to stray from the order of the topics in the topic guide if that is how the respondent wants to tell the story.

There may be occasions when an interviewer senses a respondent is feeling uncomfortable, particularly if the interview could raise sensitive issues. This is especially relevant in financial capability programs, because financial information is personal and sensitive, something we discuss in more detail at the end of this chapter. It is important to consider how such situations will be handled. This includes ways in which the interviewer will put respondents at ease before the interview begins and ways in which the interviewer will legitimize issues that arise in the interview if the respondent is feeling uneasy or is emotionally affected by the discussion during the interview. In both cases, it is important to not lead the respondent's replies in any way.

Likewise, there may be circumstances in which the interviewer may need to put respondents in touch with someone able to help with difficulties they face and to whom they should be referred. This could, for example, be appropriate if it becomes apparent that the respondent is in serious financial difficulty. In all instances, this should only be raised when the formal interview has been concluded.

Assuming there is no objection from participants, an audio recording of the interview leaves the interviewer free to concentrate on encouraging the respondent to talk freely and probe replies where appropriate and will result in a better interview as a consequence. While most respondents rapidly forget the recorder, especially if it is small and placed appropriately, it is still important to always ask for permission to use a recorder and, if participants refuse, to be prepared to take notes. The recording should be transcribed as soon as possible after the interview and reviewed by the interviewer to fill in any gaps that are inaudible to the transcriber. If the interview is in the form of interviewer notes, these should be written up as soon as is practicable. Regardless of how the interview is recorded, the interviewer should write a short note about the ambiance of the interview, including how comfortable both he and the respondent felt, as well as any distractions or interruptions, including whether there were others in the room at the time and how they could have influenced the interview.

A number of important issues must be taken into account in preparing and conducting interviews. First, the terms used to ask the questions must be familiar and understandable to the interviewee. Failure to do so may lead to adverse results by intimidating, irritating, or confusing the interviewee. Second, evaluators must make sure that questions are formulated in ways that avoid certain common pitfalls of qualitative interviewing. These pitfalls include:

- **Affectively worded questions:** This refers to questions that implicitly or explicitly reflect a value judgment about the participant such as "why did you do such a bad thing?"

- **Double-barreled questions:** This refers to questions that ask about two or more issues at the same time. For instance, "how did you use the cash provided to you through the program, and did you get any other type of assistance?"
- **Leading questions:** These are questions that may encourage the respondent to answer in a particular way. For example, "was the instruction provided by the program effective and suitable?" rather than "what was your impression of the instruction provided by the program?" The former seeds the idea that the program was effective and suitable, while the latter leaves open any possibilities—good or bad—for participant responses.
- **Ineffective question sequencing:** The order of questions makes a difference and can yield poor results, for instance, by failing to ask "easy" questions first, such as a person's household composition or economic activity, and more sensitive or complex questions later. Effective question sequencing can also help establish good rapport with the interviewer and help the interviewee feel more relaxed about the experience.

Guidance on conducting focus groups is similar to that for interviews, because we are dealing directly with participants, but it also has important differences, given the different nature of the collection method. In preparing focus groups, it is important to carefully consider their composition. This will depend on a number of questions: What kind of information does the researcher want to elicit from a focus group? How can the researcher ensure that the group composition will yield the required information in the most effective way? How useful would be it to have "control" focus groups (for instance, of nonprogram beneficiaries)?

Divisions along age, gender, social, economic, and ethnic lines may all be important factors in deciding how the groups will be formed. In certain settings, putting men and women in the same group, who are not used to speaking openly and publicly about particular issues, may lead to poor results. When programs deliver different services for different people, focus groups are most useful if they are composed of people receiving the same service. Having said that, while achieving a certain degree of homogeneity in the group is important to facilitate discussion, that does not mean that the focus group should arrive at a homogeneous view or consensus.

Moderating focus groups requires considerable skill, because a number of issues may arise. It calls for all the skills of a good in-depth interviewer, plus the ability to manage the group dynamic and ensure that everyone contributes more or less equally. The group structure means that the opinions and experiences of certain people (such as those with greater authority in their community or who are naturally more dominant) may prevail over those of others, thus providing a skewed picture.

The moderator should identify early on those who might dominate the discussion and those who might be more reluctant and ensure that the reluctant speakers are encouraged to offer their views and the dominant ones are "managed" so they contribute their views but do not monopolize the discussion.

Moderation also calls for considerable skills in steering the discussion without introducing bias into it. The probes used during the discussion should be neutral and similar to those outlined for in-depth interviews in chapter 8. It is also important to ensure that the group is stopped from being judgmental about what individual members say or do. If the discussion touches on sensitive areas, it is important to have a "cooling-off" period at the end and before the group disbands, during which the discussion moves to a more general and nonsensitive subject. As with in-depth interviews, it is advisable for a suitably skilled member of the evaluation team to facilitate at least one group. Any moderators that are not part of the team will require a full briefing to provide them with sufficient background to ensure that they capture all the points that are important for the evaluation.

It is possible, in certain instances, to obtain limited quantitative data from focus groups. For instance, focus group leaders can count the number of people who agreed or disagreed with a particular statement, or they can conduct a "ranking" exercise whereby participants rank program elements in a certain way. However, if this kind of data are required, it is extremely important to ensure that everyone's responses are recorded and that the same question is asked of all participants in each focus group. Incomplete recording of focus group participants' answers is a very common problem with this kind of method, which prevents any quantitative data from being reliable and useful. Moreover, care must be taken when interpreting such data. As focus group participants are **not** typically selected to be statistically representative of a population, these data do not give us population-wide information; rather, they are indicative of the preferences, experiences, or views of the participants alone.

Because of the nature of the discussion, an audio or video recording of the session will be necessary. The recording should be transcribed and reviewed by the moderator and any other members of the research team who participated or observed the session. And as was true for interviews, it is also important for the moderator to write a short note about the focus group as soon as possible after it has been held. This should include the dynamics of the group and any other factors that could have had a bearing on the nature and content of the discussion. This will be important when the transcript is analyzed.

In quantitative data collection, the sample size is typically much larger and conducted by a team of enumerators, who may be less individually skilled but are also required to collect less nuanced information. Many of the same interviewing

guidelines from above apply to the conduct of large-scale surveys, but evaluators also need to focus on managing the process to ensure quality and uniformity, as well as fidelity to the survey instruments. When it comes to quantitative evaluations and data collections with surveys, a detailed training and field manual for the staff should be prepared once the workplan has been developed and the survey instruments have been finalized. At this stage, quality standards should be articulated clearly and quality assurance procedures put into place before the survey begins. Without such standards in place, the survey effort can end up being poorly executed. And poor execution, in turn, can affect rates of nonresponse (refusal to answer questions) and attrition (refusal to participate further in the study), and increase the level of errors in measurement.

With such quality standards in place and clearly articulated, the next step is intensive training of enumerators and supervisors. Such training must be uniform and afford the enumerators and supervisors ample opportunity to ask questions. Usually, best practice involves a formal training program in a common location, with mock interviews and tests. This ensures not just that all interviewers can deliver the interviews and comply with the procedures and protocols, but that they do so in as similar a way as possible to avoid interviewer effects in the final data. Interviewers should also be adequately briefed about the overall study and its scientific purpose and ethical guidelines for their behavior. In addition to training and incentives, proper oversight and management of the logistics during fieldwork is an essential part of ensuring data quality and minimizing the magnitude of nonsampling errors. Whenever possible, the coordination of data collection should allow for periodic feedback from and to field teams to improve the ongoing effort.

As a general guide, most high-quality quantitative impact evaluations aim for rates of nonresponse and attrition of less than 5 percent. Explicit incentives such as compensation and bonuses for respondents and enumerators may be used to elicit such outcomes. Other procedures include frequent random quality checks at multiple levels of supervision, protocols for revisiting and verifying nonresponse units and protocols for persuading reluctant respondents to cooperate by addressing their specific concerns.

Finally, with quantitative data, data entry is typically not performed by the evaluators themselves and so during the data-entry process similar mechanisms should be instituted. Again, quality standards such as error rates should be determined in advance. Data entry operators should be trained extensively, but also provided with incentives such as bonuses and penalties tied to the error rate. Protocols to minimize transcription errors commonly include the use of a reliable data-entry software program with validation checks as well as random audits. Most critically, if data are being entered manually, double-blind data entry by two members of the team whose identity is

not known to one another is essential. Once all data are collected, all sampled units must be accounted for and summary statistics generated for review.

9.6 DATA MANAGEMENT, ACCESS, AND DOCUMENTATION

Once the data are collected, it is equally important that proper measures be taken to store data effectively and ensure that these data can be disseminated to key data users. Whoever is responsible for data collection should have hardware and software systems to warehouse data systematically and securely: Data systems, both hardware and software, are critical to effective data management and control.

Data systems should include adequate storage capacity, effective peripheral equipment to interface with data storage systems, and software compatible with program needs and the technology used. The data system should include processes to ensure that data are backed up regularly, with backup copies stored in a secure location, separate from the production data. Hardware systems should be planned to include adequate electrical power and cooling capacity to ensure reliable and continuous operation. If possible, the central elements of the data system should have redundant or replacement parts on hand if a critical component fails. These should allow for integration with the data collection systems that is as seamless as possible.

Data management goes beyond the technical issues of having the right software and hardware and ensuring backups of collected data. Part of the appropriate management of such individually identifiable data is having in place the protocols and procedures for storing, transferring, and accessing such data. All organizations that collect data on human subjects should ensure that secure storage facilities are available and used for data storage. When individually identifiable data are to be shared, data users must have proper authority to access those data, data managers must be responsible for verifying credentials of users, and data transfer systems should be secure.

All data should be stored with the minimum required individual information needed for analysis. Where possible, individual information, such as government-issued identification numbers, should be replaced by record identifiers that cannot be linked to an individual. In most cases, individually identifiable data should not be made available unless users go through a clearly defined process that verifies their need to access such data and a demonstrated capacity for appropriate data handling. A record should be retained of all users with such access.

Of course, while reliable and secure data storage is imperative, one way to make data useful is to share it with policy makers and the data-using public in "public-use" data sets. For example, your program may maintain full access to the raw data,

whereas university researchers might be given lower levels of access. However, for each class of users, maximum access should be allowed within best practices of data protection.

Final data products, primarily raw data and weights, should be accompanied by detailed information describing the variables included and their definitions, and any other information about the data that allow users to interpret and, if desired, reproduce the data collection. Data sets should be accompanied by detailed documentation, including codebooks and data dictionaries, which describe the way in which data has been coded and made available to users.

Apart from this, such documentation should also clearly describe the sampling strategy, including the design, the sample selection process, field implementation description (including nonresponse rates), any information used to inform weights, and all additional relevant information. Such documentation is both an important part of evaluation program records and an important asset for other potential future data users. Such documentation provides these users with the information they need to use the data effectively and accurately and should be provided as an accompanying document when the evaluation data are distributed.

9.7 GAINING ACCESS TO SENSITIVE FINANCIAL INFORMATION

As we discussed in chapter 8, administrative data are an important source of information for financial capability programs. For obvious reasons, access to financial records is typically more protected than many other types of data because of privacy concerns. Such data might include ATM transactions, data about credit lines or savings, and data about money transfers—barriers are to be expected in getting access to them. Yet an evaluation's objectivity and efficiency may be very much strengthened by the ability to use these types of data.

As with most other evaluations, gaining access to sensitive data involves:

- Developing and formally articulating a plan by which the necessary data can be accessed prior to launching the evaluation
- Developing mutual agreements early on with the relevant organizations and agencies for the disclosure of data, emphasizing the safeguards for privacy and security that have been put in place
- Clearly communicating the aims of the evaluation and the data and information safeguards in place to participants and stakeholders to build trust.

In some cases, compromise may be reached by allowing financial institutions to perform some analytical operations internally and provide only aggregated results to the evaluator; however, in these cases, expectations and timelines should be set well in advance, because evaluators have little control over such situations.

As noted above, when primary data are collected, participants may be reluctant to discuss or disclose information about their financial situation because of the sensitive nature of financial information. In qualitative data collection, this emphasizes the need for trained and experienced interviewers. For household surveys, in addition to piloting and testing instruments and training surveyors and moderators, other techniques for capturing data may be used. For example, one way to collect such information (from populations where literacy is high) is to give respondents privacy to fill in part of a survey themselves without interacting with the surveyor, either on a computer or in a sealed envelope to be opened later.

KEY POINTS

Once one has decided on what data collection types make sense in evaluating a financial capability program—for example a survey and set of focus groups—the process of implementing those data collection methods comes to the forefront. If such data collection is not done correctly, then it will undermine the study's validity or compromise the usefulness of the results. Proper care must be exhibited at every stage of the data collection process, which includes designing survey instruments and interview protocols, selecting the right approach to sampling, preparing and training for data collection, implementation of the instruments in the field, and finally the storage and documentation of the data collected. All of these are critical to the process of collecting data.

FURTHER READING

General

Berg, B. L. 2007. *Qualitative Research Methods for the Social Sciences.* Pearson, NY: Allyn & Bacon.

Fink, A. G., and J. Kosecof. 2008. *How to Conduct Surveys: A Step by Step Guide*, 4th ed., London: Sage Publications.

Levy, S., and Lemeshow, S. 1991. *Sampling of Populations: Methods and Applications*, New York: John Wiley & Sons, Inc.

Technical

Czaja, R., and Blair, J. 2005. *Designing Surveys, a Guide to Decisions and Procedures*, (2nd ed.), Thousand Oaks, CA: Pine Forge Pr.

Dattalo, P. 2008. *Determining Sample Size: Balancing Power, Precision, and Practicality*, New York, NY: Oxford University Press.

Levy, P. S., and Lemeshow, S. 2011. *Sampling of Populations: Methods and Applications*, 543, Wiley.

Morgan, D. L. 1997. *Focus Groups as Qualitative Research*, Thousand Oaks, CA: Sage Publications.

Patton, M. Q. 2001. *Qualitative Research & Evaluation Methods*, Thousand Oaks, CA: Sage Publications.

CHAPTER 10

Analyzing quantitative and qualitative data

A core activity in conducting evaluations—whether a process or impact one—is collecting the data needed to determine whether a financial capability program is being implemented as intended and is achieving its intended impacts. Such data—both quantitative and qualitative—will provide those answers when they are analyzed.

In this chapter, we focus on the basic concepts in analyzing quantitative data (typically for impact evaluations) and analyzing qualitative data (typically for process evaluations). More advanced readers are directed to the list of readings at the end of the chapter.

10.1 ANALYZING QUANTITATIVE DATA

Quantitative data analysis for impact evaluations involves estimating the effect of a program by comparing the outcomes of the treatment group (who received the financial capability intervention) with the outcomes of the comparison group (who did not). But what does it mean to provide a statistically valid answer to an evaluation question?

This section illustrates how to use the quantitative data collected for analysis. We will explain how to conduct a hypothesis test, the basics of regression analysis, and how you can interpret the results from the impact evaluation. The aim is to provide a critical appreciation of what is possible and credible when it comes to quantitative data analysis.

10.1.1 Descriptive statistics

Generating simple summaries helps both evaluator and eventual audiences understand what is happening in the raw data.

Once we collect any quantitative data, the first thing to do in analyzing these raw data is to generate a set of simple "descriptive statistics." As the name suggests, descriptive statistics are methods of summarizing the data, either through tabulations in tables or visually in charts and graphs. Even if the main data analysis is more complicated, which it typically will be, generating and presenting simple summaries helps both evaluators and their eventual audiences to begin to understand what is

happening in the raw data. Descriptive statistics are important in monitoring and in process and impact evaluations.

We will use the term "variable" throughout this section to be consistent with statistical terminology, but it should be clear throughout that a variable is simply a characteristic of the sample population that takes on more than one set of values (put differently, a characteristic that varies).

EXPLORING ONE VARIABLE

For monitoring and process evaluations, having a useful set of descriptive statistics that capture the status of key indicators is critical. Descriptive statistics are also an important first step in quantitative impact evaluation.

CENTRAL TENDENCY MEASURES

Mean: sample average

Mode: most frequently observed value in the sample

Median: middle score when sample is ranked from top to bottom

What do such statistics look like? In financial capability interventions, evaluators first report the sample size for the overall sample of data and then produce summary statistics of key demographic variables, such as age, gender composition, rural/urban status, education, and income (if reliable measures are available). The reason for doing so is to provide context for both evaluators and their eventual audiences and to enable some assessment of how well the sample represents the population of interest. Other variables of interest in financial capability programs that bear reporting are any measures of financial knowledge, status or behavior, such as financial literacy, total savings, or account-holding.

Such descriptive statistics can be separated into two general categories: those that measure central tendency and those that measure variability or spread. Examples of measures of central tendency are the mean (the average for the sample), mode (the most frequently observed value), and median (the 50th percentile), because these statistics are measures of the value around which the data appear to be clustered. Examples of measures of variability or spread are the range, quartile, standard deviation, and skewness. These statistics describe how the data are dispersed.

When variables are **continuous** and **numeric** (such as age in years or wealth in local currency units), we typically report the mean. However, it may also be important to consider other statistical measures that describe the distribution of the data, including the mode and the median.

Reporting mean levels alone can give a misleading picture of the sample when data are unequally distributed.

For instance, suppose that a rural bank branch serves a village in which there are a few landholders and very many landless laborers (Village A). In this village, a small fraction of the villagers are extremely wealthy, while most have nothing at all. This is quite a different setting from a village (Village B), in which most inhabitants have a moderate level of wealth. While the mean levels of wealth (the average levels of wealth) in both villages may be very similar, the mode (the most frequently observed wealth value) and the median (50th percentile of wealth when the individuals in the

villages are ranked from top to bottom) may provide a first look at the differences between the two villages. The implication from this pretty clear: Reporting mean levels alone can give a misleading picture of the sample when the **distribution of the data** is very unequal.

When variables are **categorical** (such as gender, occupation, rural/urban location), it is natural to simply report the proportion of the sample in each category (e.g., 50 percent of the sample are women). Often, categorical data can be ranked according to some criterion. When data can be ranked, we refer to it as **ordinal**, in the sense that it has a defined order. It is sometimes only possible to collect reliable data on topics such as wealth and income using ranked categories, even though the underlying information is continuous. In this case, it is useful to report the median category as well.

To continue our example, suppose we randomly sample 200 individuals from our hypothetical Village A to perform an evaluation of the bank branch's financial education program. We find that 92 percent of our sample has no land, which is then also the modal and median category, while 5 percent have land in excess of 5 acres and 3 percent have land in excess of 20 acres.

MEASURING THE SPREAD

Variance: how tightly the data are clustered around the mean

Standard deviation: the square root of the variance

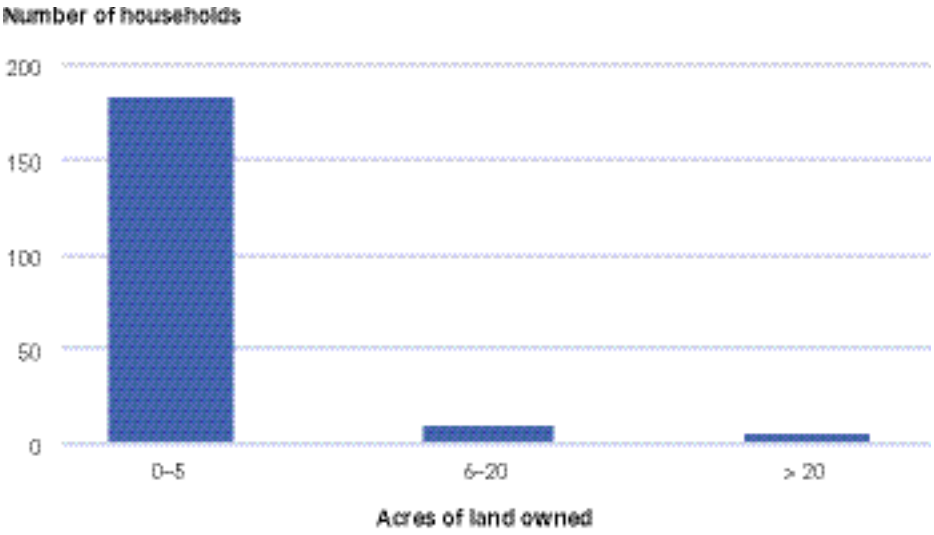
Because the measures of central tendency only provide a partial look at the data, it is important to report measures of variability as well. A simple measure of variability for ordinal data is the **range**, or the maximum and minimum values. Another set of measures is **deciles** or **quartile**. These are calculated by first sorting the sample in rank order and then reporting the scores at each 10th percentile or the 25th and the 75th percentile.

A less simple but very meaningful pair of measures of spread is the **variance** and **standard deviation**. The variance is a statistical measure of how tightly the data are clustered around the mean (technically speaking, it is the average squared deviation of each observation from the mean). The standard deviation is simply the square root of the variance. A high variance implies that the observations within the sample are very different along this measure (not clustered around the mean), while a low variance means that most are fairly similar (clustered around the mean).

In our hypothetical two villages, Village A would have a high variance for the wealth measure, whereas Village B would have a low variance for the wealth measure, even if the mean wealth was similar in both villages. While generally the range and standard deviation are sufficient to give the reader an idea of spread, one other statistic to consider is the **skewness** of the data. Skewness describes how much the distribution of values is asymmetric. Village A is an example where the wealth distribution is highly skewed, because a few people are very wealthy.

While these statistics can be reported in charts, in some cases it may be valuable to provide a **histogram** or chart that shows the distribution of the variables visually. For categorical variables the histogram would show the proportion of values in each category. Figure 10.1 shows an example histogram of land holdings in Village A as described in our example.

FIGURE 10.1 HISTOGRAM OF LAND HOLDINGS IN VILLAGE A



EXPLORING RELATIONSHIPS BETWEEN TWO OR MORE VARIABLES

We are often dealing with more than one variable and so we often want to analyze the relationship between two or more variables. This is referred to as bivariate and multivariate analyses of the descriptive statistics.

Such bivariate or multivariate analyses are key to monitoring and to process and impact evaluations. In the case of monitoring, it is important to understand if and how key process indicators change over time or vary by program location. So, for example, we may want to look at how the number of individuals trained by a program changes as the duration of the program increases, or whether there is significantly different take-up of the program in rural and urban areas. In the setting of process evaluations, we may wish to understand the relationship between certain inputs and outputs. For example, is an increase in the number of teaching hours associated with more outreach (i.e., an increase in the number of students trained), or more intensive education (i.e., more hours of training per student)? In impact evaluations, we are typically interested in the relationship between a variable that describes an outcome of interest and a variable that captures program participation.

The variable of interest can be a simple indicator of enrollment or a more refined measure, such as hours spent watching an edutainment soap opera. We typically refer to the outcome as the **dependent** variable and the participation variable as the **independent** or **explanatory** variable.

The most basic form of bivariate analysis is **cross-tabulation**. As the name suggests, this method provides a table that shows the values of one variable against the other. A commonly used extension of cross-tabulation when examining the interaction between two variables is referred to as the contingency table, which tabulates the frequencies of each combination of values in the data, or a percentage table, which expresses these as percentages.

Suppose, for instance, that the rural bank in Village A launches its financial education program and collects data on villagers' bank account ownership status over time. In our hypothetical sample of 200 villagers in Village A, we randomly allocate half of the villagers to participate in the program (giving us 100 treatment and 100 control individuals) and observe account ownership after the program.

The contingency table (table 10.1) below shows a cross-tabulation of our key outcome measure (ownership of a bank account) against our explanatory variable (participation in a financial training program).

TABLE 10.1 CONTINGENCY TABLE

	PARTICIPATED = YES	PARTICIPATED = NO
Do not have a bank account	12	56
Have a bank account	88	44

If the outcome variable is continuous (such as wealth), another way to cross-tabulate the data is to report means or other summary statistics (table 10.2).

TABLE 10.2 CONTINGENCY TABLE WITH CONTINUOUS VARIABLE

	PARTICIPATED = YES	PARTICIPATED = NO
Total savings	980 local currency units	870 local currency units

With two continuous variables, it may be more useful to use a visual representation of the relationship between the two variables by presenting a **scatterplot**. A scatterplot, as the name implies, plots the two variables on the horizontal and vertical axes

and then shows the relationship between the two by how they "scatter" in the graph (figure 10.2).

FIGURE 10.2 SCATTER PLOT OF HOUSEHOLD SAVINGS AND INCOME

The relationship between any two variables can also be captured by looking at the **correlation coefficient**. This is a standardized measure of the dependence between two variables (or covariance), which ranges from –1 to 1. A zero value implies that there is no relationship at all between the two variables. A positive value (above zero and up to 1) implies that the two variables move in the same direction (as one increases, so does the other), while a negative value (below zero and up to –1) implies that the variables move in the opposite direction (as one increases, the other decreases). A high absolute value means a stronger relationship. For instance, in the example above, the correlation coefficient between the explanatory variable (program participation), and the dependent variable (account ownership), is positive but less than 1. This tells us that participation is strongly (but not perfectly) correlated with account ownership.

Descriptive statistics uncover data patterns that help evaluators see if there are unusual trends worthy of investigation.

Descriptive statistics uncover patterns in the data and help evaluators to quickly see if there are unusual trends that should be investigated. In other words, they can be suggestive or indicative of program effects. However, they do not allow us to draw further conclusions or make predictions. Even if outcomes and participation are shown to be correlated, evaluators must do further analysis and interpretation to judge whether a program indeed had a causal effect on outcomes.

10.1.2 inferential statistics and hypothesis testing

Descriptive statistics are important to ensure that we truly understand the data, but evaluators also need **inferential statistics**, or the type of statistical analysis that allows them to draw conclusions about the causal impact of program participation on outcomes of interest. To do so, we first need to understand the concept of **statistical significance**.

TWO KINDS OF SIGNIFICANCE

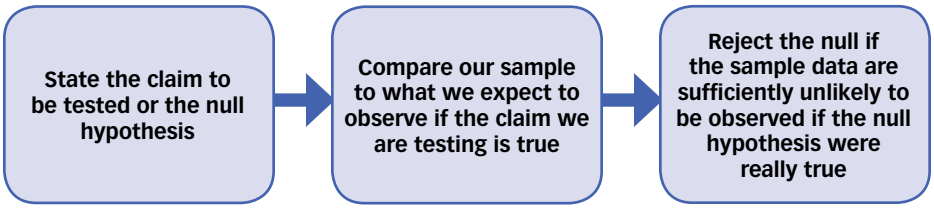
Statistical significance: likelihood that event or occurrence occurred by chance

Practical significance: the size of the event or occurrence independent of its statistical significance

The word "significance" has a very specific meaning in statistical analysis. Specifically, a statistically significant event or occurrence is something that is considered unlikely to have occurred simply by chance. Statisticians make this determination by setting up and formally testing the hypothesis or claim that the event did indeed occur by chance. This means that statistical significance is a benchmarking method that allows evaluators to precisely quantify and standardize their judgment of what qualifies as acceptable evidence for drawing conclusions. Note that the statistical significance of an event is completely separate from the **practical significance** or size of the event. An event can be small in magnitude and statistically significant or large in magnitude but statistically insignificant (or statistically not different from zero).

While statistical testing is relevant to quantitative analysis more generally, for clarity, we illustrate this concept in terms of the impact evaluation framework. The test for statistical significance occurs in three steps (illustrated in figure 10.3 and discussed below):

FIGURE 10.3 HYPOTHESIS TESTING



1. Evaluators first consider the claim or hypothesis that a program had no effect (known as **defining the null hypothesis**). Because there is always some uncertainty due to randomness, it is always possible (even though the probability may be very small) that the outcomes we observe may have occurred simply by chance.

2. Evaluators then **test this hypothesis**, by determining the likelihood that the observed patterns in the outcomes data would in fact have occurred simply by chance; in which case the null hypothesis would indeed be true.

3. Depending on this likelihood, evaluators can **reject (or fail to reject) the null hypothesis** and conclude whether the program indeed plausibly had a **statistically significant** effect on the outcomes of interest.

It may seem unusual to frame the basic impact evaluation question in terms of rejecting the null hypothesis (i.e., that the program had no effect). Conceptually, however, this directly captures the basic intuition of impact evaluations: We compare the outcomes associated with the program to the counterfactual—the outcomes of implementing no program at all—and conclude that there is a program effect only if the comparison is very dissimilar.

In our previous example, we observed that in the year after the training is completed in Village A, twice as many villagers in the treatment group had bank accounts compared to villagers in the control group (table 10.1). We also observe that treatment group participants reported somewhat more formal savings compared to participants in the control group (table 10.2). The question we want to answer is, are these differences statistically significant?

Statistical testing is done by computing a test-statistic from the sample data.

First, it is important to recognize that even without a program, there are likely to be differences in the observed behavior of any randomly drawn groups of borrowers. If the pattern we observed is very likely to happen in the normal course of things—for instance, differences of this magnitude in the level of savings are frequently observed

BOX 10.1 HYPOTHESIS TESTING: EXAMPLE FROM THE RTF PILOT PROGRAM ON SCHOOL-BASED FINANCIAL EDUCATION IN BRAZIL

The RTF pilot program in Brazil on school-based financial education examined a selection of outcomes. Below is a standard method for displaying the results, with the control group average outcome in the first column, the treatment group average outcome in the second column, and the difference (with statistically significant differences at the 5 percent level denoted with an asterisk) in the third column. There were small but statistically significant differences for almost all outcomes examined, so we may conclude the program had a broad range of effects from knowledge to behavior.

OUTCOME OF INTEREST	CONTROL (%)	TREATMENT (%)	DIFFERENCE (PERCENTAGE POINTS)
Financial proficiency test score	59	62	3*
Saving behavior: Percentage of income saved is nonzero	55	59	4*
Spending behavior: I make a list of all expenses every month	14	17	3*
Participation in household finances: Percentage of students who talk about finances with parents	70	74	4*

Note: * = statistically significant differences at the 5 percent level.

between random groups of people—the observed results cannot be taken as sufficiently strong proof that the program caused such a difference to be observed. The "effect" may well have been positive, but these differences may also have happened simply as a matter of chance. However, if it is very unlikely that such a thing would happen, we may be able to reasonably conclude that the program has indeed had an effect.

How do we go about doing a statistical test? While technical experts or analysts may actually conduct the analysis, it is always useful to understand the concepts and how the results are interpreted. In practice, a summary measure or **test-statistic** is computed from the sample data, and some standard assumptions are made about the likelihood of observing the range of values of this summary measure if the null hypothesis were true. The specific test-statistics to be used vary with the situation, but the testing procedure is the same: We compute the **p-value**, or the probability of observing a test-statistic as extreme as the sample test-statistic.

If the sample test-statistic is very extreme relative to the counterfactual (indicating that the sample is highly unlikely to have occurred by chance), the p-value will be very low. Thus, we only reject the null hypothesis if the p-value lies above a somewhat arbitrary but generally acceptable threshold of 5 or 10 percent. Statisticians refer to the cutoff threshold used for the test as the **level of significance**.

10.1.3 Regression analysis

Regression analysis is another statistical technique evaluators use to analyze how one variable is related to another variable. In a simple example of two variables, this is accomplished by fitting a straight line between the two variables, where the slope of the line is calculated to minimize the sum of the squared distance between the points and the line. The slope of this line is known as the **regression coefficient**, and it characterizes the linear relationship between the two variables. In the case of multiple independent variables, regression analysis helps explain how the typical value of the dependent variable changes when any one of the independent variables is varied, while the other independent variables are held fixed.

In an impact evaluation, regression analysis shows the change in outcomes associated with a change in program participation.

In practice, most evaluators will use regression analysis. In the case of impact evaluations, the regression coefficient gives the change in outcomes associated with a change in program participation. Regression analysis is increasingly accessible through a number of commonly available statistical packages, such as STATA, R, and SAS, as well as in some versions of Microsoft Excel. Along with these estimates, these packages will also produce test-statistics for the hypothesis that there is no program effect (that the regression coefficient on the explanatory variable measuring program participation is zero).

As noted above, if the p-value associated with the regression coefficient is **less than** 5–10 percent, this implies that the likelihood of observing the effects would be slim in the absence of the program, and we can say with high confidence that we reject the null hypothesis. In other words, we can reasonably claim that the program has a statistically significant effect. Note that this is different from what we will call **practical significance**, which tells evaluators how meaningful or important the explanatory variable is in explaining changes in the dependent variable.

Regression analysis has some important limitations to be aware of:

- It is demanding, because it requires quantitative data relating to a large number of individuals. With less data, it is more difficult to draw statistically significant results from a regression analysis.
- To draw conclusions in a regression analysis, there must be differences in the observed measures of the dependent and explanatory variables (or put differently, the variables need to have a large-enough degree of observed variance). For example, if all the observations concern the 30–40 age group, it will not be possible to estimate the influence of age on financial capability.
- The regression technique assumes that there is a linear relationship between the dependent and explanatory variables. But it is possible that there may be a strong **nonlinear** relationship between the variables that is not detected by a linear regression; fortunately, other statistical methods exist for testing this hypothesis for nonlinear relationships that can be considered.
- Outliers (a few observations with values far away from all others) can compromise the accuracy of a linear regression as they can significantly affect mean values. For example, in a financial education class, mean scores may be skewed upwards by the performance of a very small number of outstanding students. In this case, the mean and the median can be quite different! One approach is to drop the extreme outliers. While there are some statistical criteria that can be applied, often in practice judgment is required to know whether to include or exclude outliers, and which to exclude. If there is only one individual whose scores and individual circumstances are very different from the rest, the evaluator may choose to exclude him or her in order to reflect the effects on the overwhelming majority of the class. On the other hand, if there are a handful of students whose other characteristics are quite similar to the class, this usually is a call for more thorough investigation rather than dropping the observations. Regardless of the path chosen, it is important to describe all the data, including outliers, and document the choices and rationale provided before proceeding with the analysis.

Despite these limitations, in general, regression analysis is a useful and commonly applied tool to understand the relationship between two variables.

10.1.4 Estimating and interpreting program effects in practice

In chapter 6, we covered basic research designs for impact evaluations, such as randomization or propensity score matching (PSM), used in conducting a quantitative impact evaluation. In an ideal world—say for a randomized control trial (RCT) with close to full compliance, no differential attrition, and no risk of cross-contamination—the estimate of the causal intervention effect is simply the difference between the outcomes of the treatment and control groups. Thus, where randomization is planned and cleanly implemented, it is sufficient to compute and test for differences in means across the treatment and control groups.

However, as the design is implemented and data are collected, many practical challenges arise. Despite the best-laid plans, at the end of the day, these challenges can often only be overcome through careful analysis and interpretation to truly understand how much any estimated effects can be causally attributed to a financial capability program.

POTENTIAL CONFOUNDING VARIABLES

Regression analysis may show a strong link between a financial capability program and the outcomes of interest, but if other, more important confounding variables—variables that could also explain the results—have been omitted, the results in this case could be incorrect, and should be interpreted with caution.

If confounding variables are not accounted for, regression results could be incorrect.

In addition to program participation, evaluators typically include other control variables in the analysis, which are other explanatory variables that can also affect the dependent variable. Doing so controls for the many other important confounders that may cause differences in the outcomes between participants and nonparticipants. Because these confounders can occur even in a randomized trial, purely by chance, regression analysis should be used to analyze the randomized control data.

For example, you may want to control for important confounders such as background wealth in examining the relationship between account ownership and participation (because wealthier respondents may be more likely to open a bank account, all else being equal), even if the program under consideration is being evaluated in an RCT setting.

PARTIAL COMPLIANCE AND ATTRITION

In theory, all the individuals who are offered treatment in a financial capability program will participate. In reality, individuals may be offered treatment but refuse to

participate in many programs. In chapter 6, we discussed the fact that these individuals are likely to be self-selected.

PARTIAL COMPLIANCE AND ATTRITION

Partial compliance: when not all participants assigned to treatment group participate or when members in control group also receive intervention

Attrition: when participants participate but then drop out

Partial compliance occurs not only when a just fraction of the participants assigned to the treatment group actually receive the intervention, but also if some participants in the control group also receive the intervention. As discussed in chapter 6 on treatment effects, policy makers should distinguish between two kinds of "program effects":

- **Intent-to-treat estimates:** We can compare the outcomes for the group who are offered the intervention to those who are not, regardless of whether they actually received it. This is called the intent-to-treat (ITT) estimate of program effects. This estimate will tell you the average impact of the intervention for those who were targeted, because it does not adjust for intervention take-up.
- **Treatment-on-the-treated estimates:** A different statistic is the treatment-on-the-treated (TOT) estimate, which will tell you the impact of the intervention on those who received it. In other words, suppose that we want to understand the impact of the financial education program on those who actually took a financial literacy course. If participants fully comply with their treatment and control group assignments in the case of an RCT, then the ITT estimate will equal the TOT estimate, by definition.

While partial compliance may occur in many situations, in the context of financial capability program this often happens when program participation is not mandatory. Such could be the case when the program involves offering a product for voluntary purchase.

In our hypothetical example, consider that everyone in the treatment group of 100 villagers in Village A was offered the intervention, whereas no one in the control group received this offer. If all 100 villagers in the treatment group actually received the education, then we have perfect compliance and the ITT is in fact the TOT. But if only half the villagers actually received the education, then the ITT and TOT program effects may differ significantly. If we simply compare the outcomes of those offered the program to those who were not offered the program, it is important to qualify that these results reflect the ITT estimate of the program.

Instrumental variables can be applied as a solution the problem of partial compliance.

The method of **instrumental variables** can be applied as a solution to partial compliance to estimate the TOT effect of the program. (See chapter 6 for more details.) In this situation, **random assignment to the treatment group** is used as the instrumental variable. This is an indicator variable (taking a value of 0 or 1) for whether the participant was assigned to the treatment group. This variable meets the two requirements for a good instrumental variable outlined in chapter 6 because participants are more likely to take part in the intervention if they were assigned to

BOX 10.2 EXAMPLE OF PARTIAL COMPLIANCE

This study examines the impact of a financial literacy education program on the use of financial services in Indonesia. The authors Cole, Sampson, and Zia (2011) partnered with a nonprofit organization called MICRA to develop a training session explaining the use of bank accounts to individuals without such accounts. The curriculum highlighted a simple bank account, known as "SIMPEDES," that requires a low minimum deposit amount and charges no fees for four or fewer deposits or withdrawals.

A randomized control design was chosen as the evaluation methodology. To form the evaluation sample, 64 villages were chosen on the island of Java. In each village, 30 households were randomly selected, for a total of 1,920 households. To participate in the study, the households were required to have no bank account. An initial questionnaire determined that 1,173 households did not have a bank account, so these households were invited to participate in the experiment. Unfortunately, evidence showed that some surveyors were collaborating with participating households and interfering with the experiment, so a number of these households had to be excluded from the analysis. Thus, the final number of eligible units was 736, of the original 1,173 households. Of those, the evaluation sample included 564 households who chose to participate in the experiment. This is implies that 23 percent of households invited to participate in the intervention did not choose to do so. An adjustment for such partial compliance was needed.

The outcome variable of interest is whether the household opens a bank account. In this case, there was no statistically significant difference between the decision to open a bank account for treatment and control households (ITT effect). These results did not change when financial literacy program attendance was instrumented for, with assignment of a financial literacy invitation (TOT effect).

the treatment group and because being assigned to the treatment group only affects the financial capability outcomes through the act of participating in the intervention.

We should note that simply because there are analytical methods to address noncompliance, evaluators and policy makers involved in the impact evaluation must still work together to keep noncompliance to a minimum. First, because the instrumental variables method provides **Local Average Treatment Effects (LATE)**, these effects are only applicable to the participants who were either induced to take part in the intervention because of the original random assignment or encouragement in the case of instrumental variables. While in many situations this is the population of interest and may be an estimate of interest for the intervention, these effects are not generally externally valid and the findings may not be generalizable without careful consideration. Second, if the level of partial compliance is too severe and there are not enough participants remaining in the treatment group, the instrumental variable methodology will not produce correct estimates.

Another challenge evaluators face, one that is often confused with partial compliance, is **attrition**. As defined in chapter 6, attrition refers to the inability to collect

outcome data from some participants who were originally part of the analysis sample because they have dropped out of the program.

If attrition is random, it will reduce the statistical power of the evaluation (i.e., will increase standard errors), but the results will not be affected. But if the attrition is different for the treatment and control groups, it may affect the estimates of program effectiveness. For example, if more participants from the control group drop out because they are not receiving a benefit from the intervention, ignoring this fact will lead evaluators to overestimate a program's effect. If there is nonrandom attrition, then the initial randomization is compromised, and the treatment and control groups are no longer comparable. This makes attrition a very difficult problem to solve.

If attrition is nonrandom, it will compromise the initial randomization.

Thus, it is crucial to manage attrition during the data collection process. For example, it is important to collect good information about where to find participants at the outset of the study, especially if the goal is to follow participants for a long time after the end of the program (what is known as a longitudinal study). If it is not possible to follow up with all attritors, an alternative is to follow up with a random sample of the attritors. The analysis should then be adjusted to give a higher weight to those who were contacted.

When an evaluation is interested in multiple outcomes, special techniques may need to be used.

Such attrition numbers for the treatment and control groups should be reported when the results are interpreted from all analysis methods. Also, the baseline characteristics of attritors and non-attritors should be compared for both the treatment and control groups to see if there are any systematic differences. If attrition remains a problem, further statistical techniques are available to identify and adjust for the bias.

ADJUSTMENTS FOR A LARGE NUMBER OF OUTCOMES

Earlier in this chapter, we emphasized that formal hypothesis testing is necessary, because differences in the treatment and comparison groups, though unlikely, may still be observed as a matter of chance. Evaluators make judgments about statistical significance based on whether an outcome is judged to be sufficiently improbable.

With one hypothesis, this is very straightforward. But if a very large number of hypotheses are tested, we are increasingly likely to reject a false null hypothesis. In other words, although a program may have no effect, the more outcomes we examine, the more likely we are to find a difference between our treatment and control group—just by chance. For example, a researcher testing 10 independent hypotheses at a 5 percent level of significance will reject at least one of them with a probability of approximately 40 percent.

When the evaluation is interested in multiple outcomes, special techniques may need to be used when considering the statistical significance of the entire group of outcomes. The p-values must be adjusted to account for the fact that the outcome

is one of many outcomes being considered. Methods such as the Bonferroni correction (multiplying the p-value by the number of hypotheses) can be implemented with standard statistical software.

Treatment effects may not always be equal across all treated participants.

But for technical reasons, some of these methods can be extremely conservative. It is not necessary to implement a correction for multiple hypotheses with just a small set of outcomes, but it should certainly be a consideration when evaluators take on a wide range of hypotheses in order to avoid the pitfalls of "data snooping" or "data mining." In other words, this is yet another good reason for evaluators to carefully consider the selection of "key" outcomes of interest **before the start of the evaluation**.

HETEROGENEOUS EFFECTS

It is important to note that up to now we have discussed average treatment effects across all treated participants (for more discussion on treatment effects, see chapter 6). However, it may well be the case that treatment effects are not equal across all individuals or are so skewed across groups of participants that the average treatment effect does not truly represent any of the participants. This is particularly true when evaluating the equitableness of a program. For example, we may be interested in whether village financial education differently affects low- versus high-income individuals or whether the effects are different for women as opposed to men.

We may also be explicitly interested in the impact of the intervention on a specific group of participants. This is also true to understanding the appropriateness of targeting and delivery. In financial capability programs, an important factor to consider is the initial pre-intervention capability: a basic program may well have only marginal effects on the already-capable, but large effects on the less-capable. Alternatively, other, more sophisticated programs may be so advanced that it is only useful for those with sufficient fundamental skills. Cole, Sampson, and Zia (2011), for instance, find that the former scenario appears to be the case in the specific financial education program they evaluate in Indonesia: the training appears not to have an overall impact, but does work somewhat better at motivating the less-literate populations to open bank accounts.

When the population being studied is very diverse, evaluators should carefully approach treatment effects. Two common approaches that researchers use are **subsample analysis** and **interaction effects** in regressions. Subsample analysis simply means that a separate treatment effect is estimated by carrying out the analysis only on the subgroup of interest (e.g., women alone, or a certain age group alone). An interaction effect, on the other hand, is obtained using the entire sample, and estimates the difference in the treatment effect between those who received treatment who were and were not in the subgroup. In regression analysis, this is typically done by including as explanatory variables both program participation and

an additional term that reflects being in the program as well as a member of the subgroup (called the interaction term).

If possible, subgroups should be identified ahead of time, and ideally the evaluation design should account for differences in subgroups. For example, the randomization design could be stratified by the subgroups of interest, so that separate treatment and control groups are formed within subgroups. This will allow the estimation of separate treatment effects for each subgroup.

Even in cases where stratification is not possible, it is often worthwhile to explore subgroup and interaction effects on subgroups. However, if this is done ex post, two issues may arise. First, there may be insufficient power to enable analysis. Second, in relation to the problem of multiple outcomes discussed previously, as the number of subgroups analyzed increases, the probability of finding a chance effect on any one subgroup also increases.

It is important, therefore, to ensure that the analysis of heterogeneous effects is performed and justified appropriately. The best and most credible way of doing so is to generate study designs and hypotheses that involve subgroups ex ante if possible. If not, such analysis remains important and can shed additional light on the main results of the financial capability intervention, but it should be made clear that the subgroups were defined after the evaluation was in place, together with the rationale for investigating particular groups.

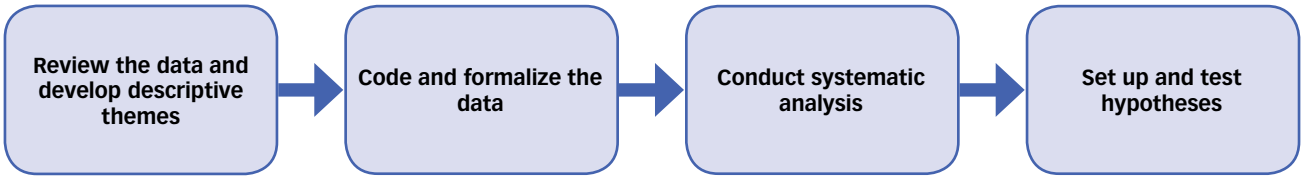
10. 2 ANALYZING QUALITATIVE DATA

While quantitative analysis is mostly done as part of impact evaluations, qualitative analysis (based largely on data and information from interviews and focus groups) is primarily used to inform process evaluations. And while the analysis of quantitative data clearly has a rigorous set of rules and procedures (as discussed above), the analysis of qualitative data in the context of process evaluations must also be done rigorously, systematically, and in a way that others can assess and replicate.

This means thinking about the analysis as the fieldwork is under way and producing analytical notes as one goes along; being systematic and deriving a structure from and for the data collected that is interpreted flexibly throughout the analysis; and aiming to go beyond the purely descriptive to provide a depth of understanding or to generate hypotheses that can be tested quantitatively. In process evaluations, the analysis of qualitative data requires that evaluators have a very thorough understanding of the program under scrutiny to ensure that findings are interpreted appropriately.

In this chapter, we discuss a rigorous approach to conducting qualitative analysis, as highlighted in figure 10.4.

FIGURE 10.4 A RIGOROUS APPROACH TO QUALITATIVE ANALYSIS



10.2.1 Review the data and develop descriptive themes

As discussed in previous chapters, the starting point for any evaluation is setting out the evaluation questions. These questions later guide the research and inform the interpretation of findings. In qualitative data collection, the initial research questions form the basis for developing **descriptive themes** along which the findings are organized. These themes can, and likely will, change as more data are collected and new patterns, issues, and lines of inquiry emerge. Still, it is important to take the initial evaluation questions as the point of departure in qualitative data analysis for evaluations.

Evaluators should develop descriptive themes to organize qualitative data.

Data analysis does not need to wait until all data has been collected. It is important to look for themes and patterns in the data even at the fieldwork stage. Thus, it is also important to keep a record of such themes and patterns as they occur to inform the plan for analysis. They can also be used to inform future data collection, adjusting the topic guide to cover areas that were not anticipated at the outset and to extend the probing of ones that emerge as being important to explore in greater detail.

Indeed, many qualitative researchers conduct their fieldwork in stages, analyzing interview or focus group transcripts as they go. If the fieldwork is being undertaken by someone other than the analyst, then debriefs will be required during which the former's insights are captured.

It is important to retain an open mind, let the key analytical themes emerge from the data, and to avoid imposing a preconceived set of themes.

Once all the fieldwork is completed and the transcripts and field notes are prepared, the first stage in the analysis is to read them all carefully—ideally more than once. This will help give a better overview and a feel for the circumstances and views of individual respondents, with the advantage of a little distance from the fieldwork itself. It is important to immerse yourself in the data, because it is this immersion that will provide you with a detailed understanding of the information and the important themes that emerge. But in doing so, it is also important to retain an open mind, let the key analytical themes emerge from the data, and to avoid imposing a preconceived set of themes. Evaluators should jot down thoughts and ideas as they go and then use these to develop an initial coding scheme, which will be mainly descriptive at this stage.

10.2.2 Code and formalize the data

The next step is to begin the coding of the transcripts or materials you have collected. Coding is a central part of all qualitative data analysis, and consists essentially of identifying the key themes in the data and labeling the data according to these themes. The themes are typically common patterns, ideas, and concerns that emerge throughout the data. As mentioned above, the process of developing these codes or themes may start as early as with the development of the evaluation questions, continue during fieldwork, and be completed and refined once all data has been collected and is reviewed.

Again, this will involve reading transcripts and materials carefully to refine, expand, or reject the initial codes. It is at this stage that evaluators will begin to move from purely descriptive codes to ones that are more analytical. They will be looking for patterns in the data: for similarities and consistencies across documents and differences and inconsistencies, with the goal of explaining why and how these occur. At the end of this process, the codes you have created can then be organized into a hierarchy of ideas (often known as a **code frame**).

10.2.3 Conduct systematic analysis

Coding the data is a means to an end, not an end in itself, and it is followed by systematizing and analyzing the information collected. There are a number of ways of doing this, and individual researchers each have their preferred method. Some prefer to work with annotated transcripts, some use computer packages, and others use thematic grids. Each of these is described briefly below.

WAYS OF DOING SYSTEMATIC ANALYSIS

Folders: organize issues from transcripts and other documents into folders, either manually or with computer programs

Thematic grids: summarize indexed transcripts or other documents onto table grids according to hierarchical code frames

Some qualitative researchers create a series of **folders**, one for each issue or theme, and assign the sections of text from transcripts (with an identifier to link it back to the full script and, in the case of interview transcripts, the characteristics of the respondent) to them. It is not uncommon for sections of text from transcripts to be relevant to more than one theme, which means it is good practice to include these sections under all the codes they address. Having assigned all the material from the transcripts to the different folders, it is then possible to work through them one by one to bring the information in them together, analytically. This approach is very flexible and allows you to modify, combine, or split categories as the analysis proceeds and new insights emerge.

Computer programs help emulate this process and have become more sophisticated over the years. Many now permit the integration of data in different formats, such as interview transcripts, web pages, audio or video recordings, or extracts from social media sites. Still, the process is essentially the same as the folders process described above, creating an initial, largely descriptive, electronic coding frame initially and

then refining that into one that is both analytical and hierarchical. But it is important to remember that computer programs of this kind do not analyze the data for the evaluators; they merely enable evaluators to store, organize, search, categorize, and group large volumes of mostly text-based data. Examples of computer software for the analysis of qualitative data are NVivo, ATLAS.ti and Dedoose, among others.

The advantage of using a computer program is that "interrogating" the data becomes far easier and much quicker. Some even have automatic indexing for highly structured data, such as semi-structured interview scripts where identical questions are asked in a predetermined order. Computer programs are particularly valuable where there is a large amount of data to be analyzed. But researchers who prefer the manual approach argue that it enables them to retain a more holistic view of the data than they feel is possible with a computer program, and that the manual approach is more suitable for in-depth analysis of small numbers of transcripts.

Other researchers prefer to systematize their data using **thematic grids**. Having indexed the transcripts or other documents, evaluators then summarize them on grids, with the columns corresponding to the hierarchical code frame developed and transcripts or documents entered in the rows. In preparing the grids, evaluators generally take the analysis a step further than the coding. Advocates of this approach claim that it makes patterns in the data easier to see and may highlight linkages that might not be apparent from the previous two methods. It also retains a holistic view of individual respondents or focus groups. Then again, it does take longer to systematize the data using the thematic grid approach.

Thus, the choice of which method to use is partly a personal one, but it also depends on the number and nature of the documents to be analyzed. Again, though, none of the three approaches is an end in itself, unless the aim is to write a purely descriptive account of the data.

Once the data have been organized into folders or grids, the researcher then has to look for patterns and explanations of those patterns, keeping a detailed log of the analysis as it goes. A good qualitative evaluator will test and challenge the interpretations as they evolve and will triangulate between different data sources.

A good qualitative evaluator will interpret the data but will also remain true to what participants said and how they said it. They will also maintain the participants' voice when analyzing transcripts of interviews or focus groups, using direct quotes in the report that they write. Above all, it is important to realize that there are no shortcuts to rigorous qualitative analysis. It is time-intensive work.

KEY POINTS

Assuming that the data for evaluation have been collected using the appropriate quantitative and qualitative tools, the analysis of the data will inform evaluators and other stakeholders of whether a financial capability program is being appropriately and effectively fielded and whether participants are improving in terms of the outcomes of interest that the program was designed to address.

But the credibility of those results depends on how well evaluators conduct the analysis of the data collected. Whether evaluators are analyzing quantitative data collected during experimental or nonexperimental research designs or qualitative data from interviews, document reviews, or focus groups, they must approach these analyses rigorously and systematically, following the procedures and rules that govern the types of analysis.

FURTHER READING

General

Maxwell, J. A. 2004. *Qualitative Research Design: An Interactive Approach*, Thousand Oaks, CA: Sage.

Robson, C. 2002. *Real World Research: A Resource for Social Scientists and Practitioner-Researchers*, vol. 2, Oxford: Blackwell.

Technical

Caracelli, V. J., and Greene, J. C. 1993. "Data Analysis Strategies for Mixed-Method Evaluation Designs," *Educational Evaluation and Policy Analysis* 15 (2): 195–207.

Caudle, S. 2004. "Qualitative Data Analysis," *Handbook of Practical Program Evaluation*, vol. 19, 417.

Cole, S., T. Sampson, and B. Zia. 2011. "Prices or Knowledge? What Drives Demand for Financial Services in Emerging Markets?" *Journal of Finance* 66 (6): 1933–67.

Duflo, E., R. Glennerster, and M. Kremer. 2007. "Using Randomization in Development Economics Research: A Toolkit," *Handbook of Development Economics*, vol. 4, 3895–962.

Grbich, C. 2007. *Qualitative Data Analysis*. Trowbridge, Wiltshire: The Cromwell Press Ltd.

Khandker, S. R., G. B. Koolwal, and H. Samad. 2009. "Handbook on Quantitative Methods of Program Evaluation," Washington, DC: World Bank.

Miles, M. B., and Huberman, A. M. 1994. *Qualitative Data Analysis: An Expanded Sourcebook*. Thousand Oaks, CA: Sage.

Shaffer, J. P. 1995. "Multiple Hypothesis Testing," *Annual Review of Psychology* 46 (1): 561–84.

PART IV

Other issues in conducting M&E for financial capability programs

CHAPTER 11

Cost analysis: weighing program costs and benefits

Laying out the economic case for a financial capability program is an important part of evaluation, even more so when resources may be scarce. Evaluators may demonstrate through impact analysis that the program did indeed achieve some or all of the outcomes set out, but that does not necessarily mean it was worthwhile to invest in the program. Some kind of judgment is needed about whether it is worth investing in a specific program or intervention and that judgment generally revolves around answering two questions for policy makers and other stakeholders:

- Did the benefits of the program outweigh the costs of undertaking it?
- Did the program perform efficiently when compared to other options addressing the same program goals and objectives?

For instance, a financial capability program may have large benefits, but it may also be extremely costly to implement. If the costs outweigh the benefits, then the program is likely to be stopped. Then again, even if the benefits do outweigh the costs, there may be other less-costly ways of getting the same or similar benefits; in such a case, policy makers may determine that the program—good as it is—may not be the best way to use the limited resources available. And, of course, the choices policy makers face may not be so clear-cut: It could turn out that a different program may have fewer benefits than the one in question, but also cost significantly less.

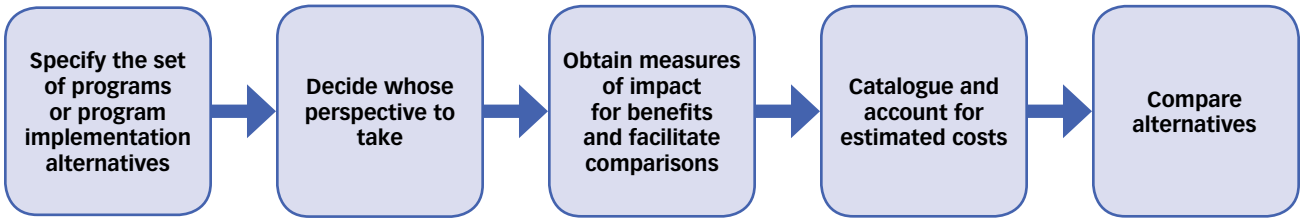
To help address these important decisions, evaluators can conduct cost analyses that systematically catalogue the benefits and costs of the program. Note that doing so requires having the benefits part of the equation, which means that such analysis can only take place after the program and its impact evaluation—which show what benefits occurred—is completed. In this chapter, we discuss the steps in cost analysis and provide simple illustrations based on actual examples. Cost analysis is a large subject, and the Toolkit covers only these basic principles. The more advanced reader is directed to the resources listed in the References.

Cost analyses can only take place after an impact analysis has been done.

11.1 WHAT ARE THE STEPS IN CONDUCTING A COST ANALYSIS?

Figure 11.1 shows the basic steps to conduct any cost analysis. Evaluators should start by specifying the set of programs or program implementation alternatives. That is followed by deciding whose perspective to take in the cost analysis; obtaining the measures of impact for benefits and facilitating comparisons; cataloging and accounting for estimated costs; and, finally, comparing alternatives. The steps are discussed in more detail below.

FIGURE 11.1 BASIC STEPS FOR COST ANALYSIS



11.2 SPECIFYING PROGRAM ALTERNATIVES

In any cost analysis, it makes sense to first start by defining the alternatives under consideration. In many instances, it may be obvious what the alternatives under consideration are: We are comparing two programs, or the value proposition of a program to no program. However, in some cases, a program may have any number of dimensions and variations, and these variations will most likely change the program's impact and costs.

For example, a school-based financial education program might have a number of different curricula that can be taught to children of different ages. However, in practice, it is only feasible to analyze a few of those alternatives. Because there could be any number of possible dimensions along which a program might be changed—and therefore a large number of alternatives to consider—it is important to first restrict the attention to a very limited number that have high chances for success in achieving the desired policy goal.

11.3 CHOOSING A PERSPECTIVE (OR SEVERAL)

The next step is to recognize that a program's costs and benefits are realized by many different stakeholders (participants, the program itself, the government, and society as a whole). In other words, in determining costs and benefits of a financial capability program, the perspective matters.

In determining the costs and benefits of a program, the perspective matters.

In our bank example, suppose local clients attending the program spend a significant amount on transportation and forgo time spent working on their own businesses to attend the training sessions. These costs are important to the clients (who lose time and money) and to society as a whole (which loses productivity), but those costs may not be relevant when evaluating the program's business case—the program's underlying reason for being conducted.

Alternatively, if a local government is financing and implementing a financial capability program, it might only take into account the local costs and benefits of the program. In reality, however, local interventions, such as educational campaigns to encourage individuals to stay out of debt, might have a broader geographic impact. The cost-effectiveness of a program might change depending on whether more global costs and benefits are counted.

Thus, in conducting a cost analysis, the evaluator must define the stakeholder (or stakeholders) from whose perspective the analysis is being conducted before he or she is able to identify the impacts of relevance and catalog the relevant costs. If multiple stakeholders are of interest, it may be relevant to do a separate cost analysis for each.

11.4 MEASURING BENEFITS AND COSTS

Like any other form of evaluation, the value of a cost analysis relies on what goes into it—in this case, on the accuracy of the estimated benefits and costs. The previous chapters of the Toolkit cover methods used to select measurement indicators that reflect the benefits chosen. Here we simply reiterate that it is critically important for evaluators to consider only those impacts that are directly attributable to a program to accurately reflect the trade-off between the program's costs and benefits.

The value of a cost analysis relies on the accuracy of the estimated benefits and costs that go into it.

It may seem that the process of estimating benefits is far more complicated than that of estimating costs, especially if programs are able to report their total budgets. However, getting at the true **economic** costs of a program can be more complex than it first appears. At first glance, it may seem that a program's budget captures its costs fairly well. However, this may only be partially true.

Some of those costs may be straightforward. For example, in the case of a high-school financial education program, such straightforward costs will include the price of the teaching materials and the wages of the teachers. But costs include more than just dedicated project spending. For instance, they also include the value of inputs that are funded by partners and those that are donated or volunteered (e.g., building space to conduct the program or computers). These inputs are also costly to someone in the sense that they could have been used for some other purpose (e.g., a person could have worked for wages instead of volunteering).

To understand the importance of accounting for all the costs, we need to again consider the concept of the counterfactual that we discussed in the chapter on impact analysis (chapter 6). If the program were not in place, resources used would be deployed elsewhere. Thus, from an economic perspective, the relevant benchmark for cost comparisons is the **opportunity cost**, or the value of the best possible alternative use of all program resources. Accounting budgets may or may not capture opportunity costs adequately, especially when resources are obtained at other than market prices or shared with other programs. In this sense, "costs" are different from "expenditures," which refer to money spent.

In accounting budgets, costs are different from expenditures, which refer to money spent.

For instance, suppose that when a financial education program is held at a bank, the bank pays for the costs of program materials and for renting a training facility as part of its program budget. However, no more money is left in the budget for professional trainers, so a few salaried bank staff are asked to conduct the program as part of their regular workload. While there is no additional expenditure to the program, the true cost to the bank actually includes the value of its time (i.e., profits forgone, because those members of its staff would otherwise have been conducting bank business). Ignoring opportunity costs in the example would lead to inaccurate cost estimates.

To ensure that you are thinking about costs in a comprehensive way, some evaluators recommend thinking about a checklist of important categories of program inputs. These key categories and some examples are illustrated in figure 11.2.

Finally, it is important to account for the nature and timing of different program costs vis-à-vis benefits. Financial tracking systems often distinguish between recurring and capital costs, as well as actual and planned costs. While the time period for recurring costs is generally obvious, calculating the costs of capital (such as computers and cars) is more difficult because these inputs are used repeatedly for a long period of time. Often, your government or international organization can provide depreciation tables for the most common capital inputs. You may wish to separately identify costs associated with the initial project startup from the ongoing operating costs.

It is important to account for the nature and timing of different program costs vis-à-vis the benefits.

Further adjustments must be made if programs last several years:

FIGURE 11.2 KEY COST CATEGORIES

Labor	Staff costs (market wage value + benefits)
Land	Facility costs
Capital	Equipment
Participant inputs	Transportation costs, forgone income
Other	Supplies, utilities, and other costs

- If the cost-benefit analysis is prospective (i.e., looking out to the future), projections of costs and benefits need to be made that reflect implementation. For instance, if a program consists of a six-month course followed by periodic follow-up meetings every month over two years, then the projected program costs will vary significantly between Year 1 and all subsequent years.
- Since most people prefer to consume now rather than later, evaluators should also be aware that, in practice, one dollar now has a higher value than one dollar next year and ensure that any future benefits and costs (projected or actual) are discounted appropriately (i.e., reduced by an appropriate interest rate).

11.5 COMPARING ALTERNATIVES

Here, we turn to a discussion of the different kinds of cost analyses that are typically conducted and what they entail.

11.5.1 Cost-benefit analysis

In a CBA, all impacts are monetized, including intangibles.

In a cost-benefit analysis (CBA), all impacts are monetized, including intangibles. To do this, evaluators need estimates of the different returns to financial market participation, such as access to credit, different savings mechanisms, and financial security. Ideally, these estimates should be specific to the geographical area and country where the program is being implemented. For example, in the case of a financial education program intended to reduce the fraction of unbanked individuals in a target population, a monetary value is assigned for each extra account. This is easier for some outcomes stakeholders than others: for instance, a bank may be easily able to place a monetary value on an additional customer, while an individual may not be

able to equate his/her personal financial security with a dollar value. However, it may be possible to estimate the additional monetary benefit of having a bank account in terms of the interest gained on average balances compared to holding the money in cash.

The crux of a CBA is to compare the monetary value of the benefits to the costs for a given program. Evaluators do this by computing a number of alternative measures, such as:

- **Net Present Value (NPV)** represents the value of benefits minus costs. If the costs exceed the benefits, the NPV will be negative.
- **Return on Investment (ROI)** is the NPV of the project divided by the total costs. If the costs exceed the benefits, ROI will be negative.

These measures allow policy makers to rank multiple programs in terms of their economic value. For a single program, decisions can be made about whether to continue the program by comparing the program's NPV or ROI to a threshold value.

For instance, funders may decide that as long as the program has benefits in excess of costs, it should be continued. Other organizations may need to consider whether the programs' ROI compares well to a predetermined target.

11.5.2 Cost-effectiveness analysis

In a cost-effectiveness analysis (CEA), multiple programs are compared in terms of their impact on a single, common outcome, which may be expressed in monetary terms or in other units. In such analyses, evaluators can compute **cost-benefit ratios**, or the monetary value of costs over benefits for a single outcome

In a CEA, multiple programs are compared in terms of their impact on a single, common outcome.

For instance, we may wish to compare two programs aimed at reducing the unbanked rate in a rural population. In a CEA, we would wish to know the cost per bank account opened in Program A versus Program B. Alternatively, we may express this as the inverse, the **benefit-cost ratio**, or the number of bank accounts opened per dollar in Program A versus Program B. Note that unlike a CBA, a CEA is inherently comparative: The question of whether a program is cost-effective can only be answered **relative** to another program.

In the case of programs with multiple outcomes, it is important that the CEA be conducted on a chosen measure that captures the key program objective. An alternative way to conduct a CEA is to generate a single index by attaching specific weights to the outcome measures of interest and then computing the costs needed for a unit increase in the index. However, this can be less straightforward and more difficult to interpret, especially if stakeholders do not reach a consensus on appropriate relative weights of the outcomes. For example, suppose a program run by a

BOX 11.1 FINANCIAL EDUCATION VERSUS SUBSIDIES AND MONEY-BACK GUARANTEES IN INDONESIA AND INDIA

Cost-effectiveness analysis does not have to be complicated to be useful. Cole, Sampson, and Zia (2011) examine whether a financial education program in Indonesia leads to an increase in the use of bank accounts. They compare this intervention to another program that gives small subsidies for opening a bank account and find that these subsidies increase demand more.

The total literacy training cost is approximately $17 per program participant to deliver. Among those with low levels of initial financial literacy, the training program increased the share who had a bank savings account by approximately 5 percentage points. Thus, the opening of one bank account costs $17/0.05 = $340.

In contrast, for this same subsample, providing a subsidy of $11 led to a 7.6 percentage point increase in the probability of opening a savings account, suggesting a cost per bank savings account opened of $11/0.076 = $145.

Thus, subsidies are almost two and one-half times more cost-effective than the financial literacy education program based on direct costs alone. While this calculation does not take into account any of the other costs related to the two programs, in this context this provides policy makers with sufficient information to conclude that the subsidies are more cost-efficient, given that additional costs are likely to be similar or higher for the training program.

bank but supported by external funders aims to improve the general financial capability of consumers as measured by their bank account ownership and their general financial knowledge. The bank may wish to focus attention on account ownership behavior, while the funders may wish to judge the program more on changes in general financial knowledge.

Another approach to standardization across programs and outcomes is to convert outcome measures to a common utility score or monetary value (as in a CBA). For instance, it may be possible to ask stakeholders how much they would be willing to pay for various outcomes or to ask stakeholders to assign utility rankings to such outcomes and then impute the total value of benefits either in dollar or utility terms for the computation of cost-benefit ratios.

When utility scores are used, this is referred to as **cost-utility analysis**. The advantage of this, in theory, is that it allows evaluators to directly compare the welfare benefits of one program against another. In practice, however, obtaining the data needed to perform this conversion is not trivial, and cost-utility analysis is seldom used in financial capability interventions.

One thing that should be very clear from the above discussion is that conducting CBA/CEA often relies on making a large number of assumptions. On the cost side, for instance, we may not know the wage rates for volunteer labor but still have to make an assumption about the prevailing market wage. Or, we may need to make an assumption about appropriate discount or depreciation rates (if the program is long-term).

On the benefit side, our estimates of benefits may be statistically determined to lie within a broad range of possibilities. Thus, many evaluators often find it wise to compute their main cost analysis with the most reasonable assumptions but to also conduct and report different versions using different assumptions to give the reader an idea of the results' **sensitivity**.

A CCA complements a CEA or CBA and helps policy makers understand the bigger picture of costs and benefits.

11.5.3 Cost-consequences analysis

An important alternative to a CEA/CBA is cost-consequences analysis (CCA). A CCA is simply a table or list that enumerates and characterizes all relevant costs and benefits for the alternative programs, side by side, numerically where possible and qualitatively where not. Where there is neither rigorous qualitative nor quantitative evidence, the CCA should indicate that no evidence is available. A CCA provides a way to visually integrate the qualitative and quantitative data collected from an evaluation. Tabulating costs and benefits in a CCA is an easy way to help policy makers understand the bigger picture of costs and benefits. For instance, a simple cost-consequences table that compares alternative financial education programs aimed at increasing the number of bank accounts in a rural village might look like the example in table 11.1.

TABLE 11.1 EXAMPLE OF A CCA THAT COMPARES ALTERNATIVE FINANCIAL CAPABILITY PROGRAMS

		ALTERNATIVE FINANCIAL CAPABILITY PROGRAMS		
		A	B	C
BENEFITS	Likelihood of opening a bank account after training	50%	30%	10%
	Social acceptability	Very high	High	Low
COSTS	Direct costs (labor, materials, and other) per individual trained	$10	$5	$2
	Indirect costs (facilities, overhead costs etc.)	Not evaluated	Not evaluated	Not evaluated
	Time cost for participants	0.5 hours	5 hours	10 hours

CCA ultimately is descriptive. It leaves the judgment of what costs and benefits should be included to the decision maker. Wherever circumstances permit, of course, a robust numerical CEA/CBA should be conducted. However, if a robust quantitative cost analysis cannot be performed, a comprehensive and transparent CCA analysis may be more credible and ultimately more useful for decision makers—especially if compared to a potentially misleading cost-effectiveness analysis with weak or questionable assumptions.

Table 11.2 captures the differences between the three types of cost analyses.

TABLE 11.2 TYPES OF COST ANALYSES

	COST-BENEFIT ANALYSIS	COST-EFFECTIVENESS ANALYSIS	COST-CONSEQUENCES ANALYSIS
What are the alternatives for comparison?	One or more programs to a threshold or benchmark for decision making	Two or more programs	One or more programs to one another
How many outcomes?	Multiple	One	Multiple
How do we make comparisons?	Convert all outcomes to a dollar value; then compare costs per dollar benefits	Select a common outcome measure for each intervention and compare costs per unit outcome across all programs	List and characterize all costs and benefits across all programs (qualitatively and/or quantitatively)
Result	Ratio of costs to benefits, "return on investment," or "net present value"	Ratio of cost to common outcome measure	Costs and benefits displayed in tabular form

While cost analysis is important and useful, it can also be difficult when a program is complex. At this time, few evaluations in the financial capability literature include an explicit quantitative cost analysis, partly because some forms of cost analyses are so challenging. In practice, evaluators frequently cannot abstain altogether from any discussion of cost. In the end, the form of cost analysis that is most appropriate depends on the question at hand and the quality of the relevant available data.

KEY POINTS

While evaluation often focuses on whether a financial capability program has achieved the program's objectives and has had an impact for specific outcomes in terms of the indicators of interest, the impact or benefit of a program does not occur in a vacuum. There are costs to implement any program, and if policy makers and stakeholders want to understand whether the investment in a particular financial capability program was worth it, then costs analyses will be necessary.

There are a number of ways to conduct cost analyses—including cost-benefit analyses, cost-effectiveness analyses, and cost-consequences analyses—but the key is to make sure to estimate all the costs and benefits accurately, both the straightforward and tangible ones and the less straightforward and intangible ones. Not considering all the costs and benefits when doing a cost analysis undermines the usefulness of the results.

To understand whether the investment in a particular financial capability program was worth it, costs analyses will be necessary.

FURTHER READING

General

Cole, S., T. Sampson, and B. Zia. 2011. "Prices or Knowledge? What Drives Demand for Financial Services in Emerging Markets?" *Journal of Finance* 66 (6): 1933–67.

Independent Evaluation Group, World Bank. 2010. *Cost-Benefit Analysis in World Bank Projects*. Washington, DC: World Bank.

McEwan, P. J. 2012. "Cost-effectiveness analysis of education and health interventions in developing countries," *Journal of Development Effectiveness* 4 (2): 189–213.

WHO-CHOICE. 2013. *Choosing Interventions That Are Cost-Effective*. As of February 13th 2013: http://www.who.int/choice/en.

Technical

Adam, T., Evans, D. B., and Koopmanschap, M. A. 2003. "Cost-Effectiveness Analysis: Can We Reduce Variability in Costing Methods?" *International Journal of Technology Assessment in Health Care* 19 (2): 407–20.

Baltussen, R., Ament, A., and Leidl, R. 1996. "Making Cost Assessments Based on RCTs More Useful to Decision-Makers," *Health Policy* 37 (3): 163–83.

Hutubessy, R., Chisholm, D., and Edejer, T. T. 2003. "Generalized Cost-Effectiveness Analysis for National-Level Priority-Setting in the Health Sector," *Cost Effectiveness and Resource Allocation* 1 (1): 8.

Murray, C. J., Evans, D. B., Acharya, A., and Baltussen, R. M. 2000. "Development of WHO Guidelines on Generalized Cost-Effectiveness Analysis," *Health Economics* 9 (3): 235–51.

CHAPTER 12

Implementing the evaluation

Earlier chapters have reviewed the conceptual foundation for conducting an evaluation, both in terms of developing an understanding of the program at hand and the principles, designs, and methods of monitoring and evaluation, using illustrations from the Russia Financial Literacy and Education Trust Fund (RTF) pilot programs.

This chapter discusses the practical aspects of translating these building blocks into an actionable plan for conducting an evaluation. It is important to note that in this chapter, many examples are discussed in the context of a prospective evaluation, but are relevant to evaluation activities in general. The steps discussed in this chapter are not necessarily sequential.

12.1 LOGISTICS AND TIMING

Four key factors that should be considered for the logistics and timing of the evaluation are the program cycle, the reasonably expected time needed for results to manifest themselves, the logistical constraints of fieldwork, and internal or external decision points that affect the program. Figure 12.1 shows an example of prospective evaluation relative to the main components of a simplified program cycle. Just as the evaluation design should be fitted to the program design, evaluation timing also needs to be fitted to the program cycle.

Just as the evaluation design should be fitted to the program design, evaluation timing also needs to be fitted to the program cycle.

During startup, baseline data are collected to establish the initial conditions. During execution, monitoring data are collected to track the use of inputs, the conduct of activities, and the production of outputs. After execution (or during a midline survey), measurement of short- and long-term outcomes can be conducted and compared to the baseline. Feedback loops can then take place between program and evaluation during startup, during the course of monitoring, after midline, and/or after the final evaluation.

Incorporating such measures as a parallel system greatly enhances the quality of evaluation for several reasons. In addition to the benefits of providing immediate feedback to the program itself, planning from the start can improve the collection of baseline data (ideally longitudinal) for comparison purposes, increasing the chance

that results are credible and that evaluation is viewed as ongoing and constructive. Building a parallel system may also result in efficiencies from a cost perspective, because the activities for monitoring and evaluation can then be structured to take advantage of program infrastructure in a manner appropriate to the scale and resources of the program.

The collection of baseline data should predate any program-related activity. If there is significant concern about lasting interview effects, some amount of time should elapse before the intervention itself.

The ideal timing of the follow-up data depends very much on the nature and time frame for expected results and the number of follow-ups that can be accommodated. Following up too early can result in only partially capturing long-term outcomes, simply because effects have not yet been realized.

Alternatively, in some cases, immediate follow-up may be viewed as less convincing, especially if no other long-term follow-up is scheduled. For instance, in the case of school-based financial education, it is reasonable to measure immediate impact on test scores and intended behaviors, but it is unreasonable to expect immediate changes in significant financial decisions. Following up too late, on the other hand, can be logistically complicated, and may result in the failure to capture important short-term program impacts that may decline over time.

The logistical demands of fieldwork and other background factors may also affect timing.

The logistical demands of fieldwork and other background factors may also affect timing. The logistical plan should allow enough time for travel, including possible delays between destinations that may be far-flung or occasionally inaccessible. The schedule should also allow for rest periods for field staff and account for important seasonal events, including festivals, planting and harvest, and rainy/dry seasons, which affect both staff and respondents' willingness and availability to participate.

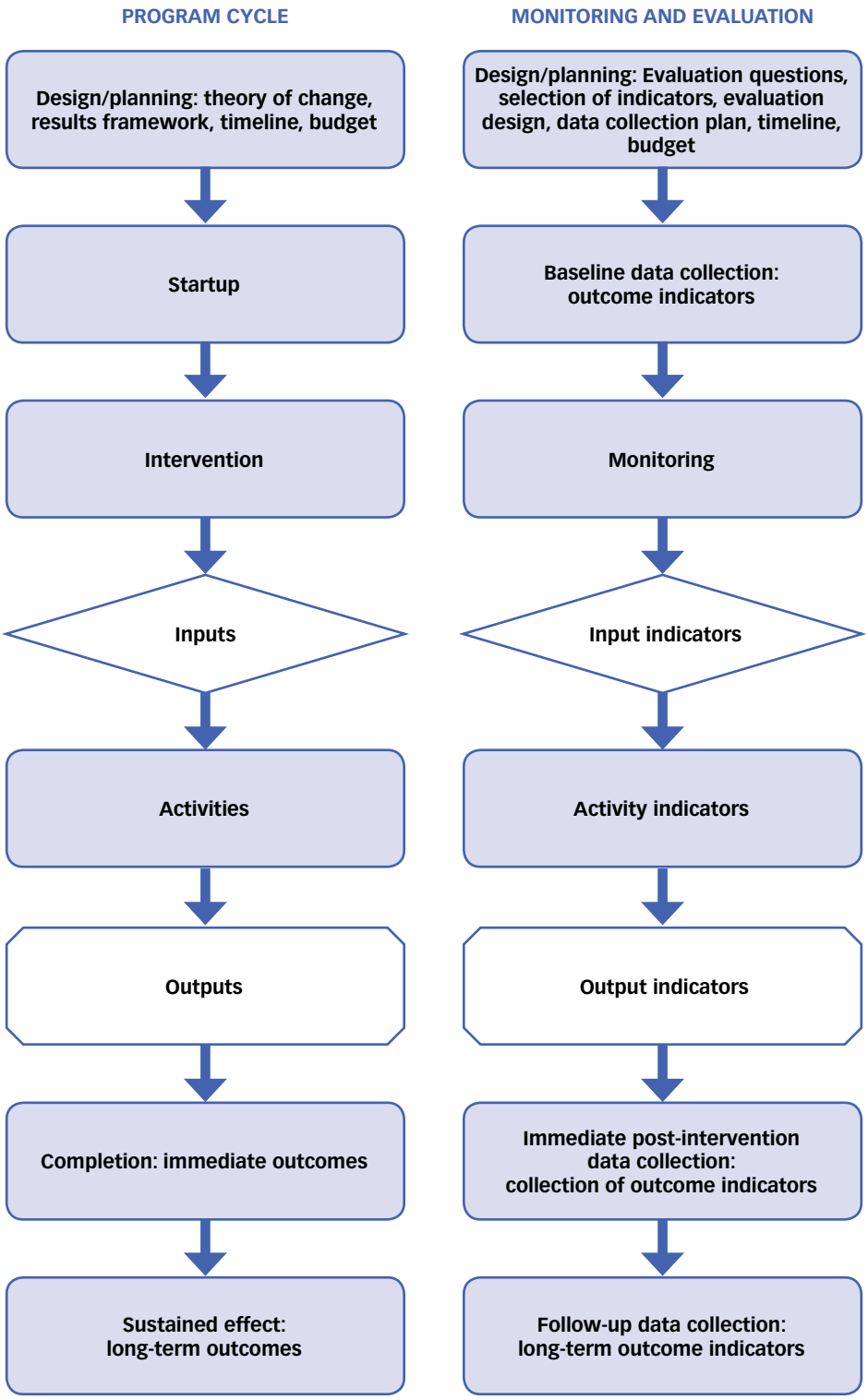
Finally, evaluations may often face external constraints on timing resulting from the need to provide input into policy making. For instance, a five-year financial capability program may face funding renewals after two years; as a result, the program may require evidence of its performance, regardless of the conditions that prevail in the field.

12.2 FORMING AN EVALUATION TEAM

Individual team members may play multiple roles (i.e., "wear multiple hats") based on capacity and resources.

Once the general scope of the evaluation is determined, it is possible to consider the range of skills and roles needed to design it and carry it out. In most instances, an evaluation team will be needed to meet the evaluation's diverse needs. Selecting the team is critically important, because the **quality**, **timeliness**, **objectivity**, and **usefulness** of the evaluation depend greatly on the chosen team's capacity and

FIGURE 12.1 MEASURING PERFORMANCE AND EFFECTIVENESS: EXAMPLE OF LINKING TO THE PROGRAM CYCLE



its members' ability to effectively work together. The process of staffing the team; ensuring a clear and mutually agreed–upon set of roles and responsibilities; and setting up a system for discussion, ongoing communication, and feedback can be critical to evaluation success.

Individual team members may play multiple roles (i.e., "wear multiple hats") based on capacity and resources. Indeed, although many evaluations are carried out by teams, in practice team members often take on several roles and responsibilities. In such cases it is helpful to combine roles that have overlapping tasks but ensure that each person can receive support and feedback from other team members. Table 12.1 lists the skills and roles that should be represented in the core team.

TABLE 12.1 CORE EVALUATION TEAM ROLES AND RESPONSIBILITIES/SKILLS

ROLE	RESPONSIBILITY/SKILLS
Evaluation manager	Overall responsibility for establishing the information needs and indicators for the evaluation, drafting terms of reference for the evaluation, selecting the evaluation methodology, identifying the evaluation team, oversight of tasks, and report writing.
Analyst	Determines evaluation design, together with the evaluation manager. A quantitative analyst conducts the quantitative analysis and participates in report writing. A qualitative analyst ensures participatory input and qualitative analysis at different stages of the impact evaluation. The analyst may be trained as an economist, sociologist, anthropologist, or other social science analyst depending on nature of evaluation.
Information coordinator	Gathers and coordinates information or data from various sources

If data are to be collected from participants, technical experts and a fieldwork team will also be needed, as shown in table 12.2.

TABLE 12.2 DATA COLLECTION EVALUATION TEAM ROLES AND RESPONSIBILITIES

ROLE	RESPONSIBILITY
Statistician or sampling expert	Responsibility for selecting sites and groups for pilot testing, site and sample frame and sample selection, and sample size calculation and generation of sampling weights if needed for quantitative analysis.
Survey designer	Responsible for designing the data collection instruments and accompanying manuals and codebooks so they are useful, context-appropriate, and feasible for implementation, as well as for pilot testing and refinement of questionnaires.
Fieldwork manager	Responsible for overseeing the data collection effort
Field team—interviewers and supervisors	Responsible for delivering surveys and conducting other data collection efforts
Data entry team—data entry operators and supervisors	Responsible for conducting data entry and quality control procedures

Important decisions must also be made about the allocation of internal and external responsibility. In practice, an evaluation team often needs to strike an important balance. Using external evaluators can help achieve the right mix of technical skills and establish the objectivity and credibility of evaluation. But there are important reasons to keep internal stakeholders engaged:

- Keeping the budget manageable by using in-house staff.
- Building capacity and awareness within the organization.
- Ensuring buy-in and recognition of the evaluation process such that the logistics are smooth (particularly if randomization is used) and that the results are put to use to inform policy.

In practice, an evaluation team needs to strike an important balance between using external and internal evaluators.

Thus, contracting with an external evaluation manager or analyst, if budget permits, can be desirable in many larger-scale evaluations, although internal stakeholders may have an advisory role. A decision should be made whether to use an established team or to select and work with a group of individual evaluators. An established team may be more effective and efficient, but it may be more costly; then again, coordinating individual consultants may allow for the best mix of qualifications, but it may also impose a large burden on the coordinating manager. In either scenario, when engaging external evaluation consultants it is important to allow for significant time to plan and craft the terms of reference (ToR). The ToR establishes the scope of work and includes objectives, key tasks, methodology, timeline, and key deliverables.

BOX 12.1 THE OPPORTUNITIES OF WORKING WITH A HIRED SURVEY COMPANY: EXAMPLE FROM AN RTF PILOT

In the example RTF pilot program in South Africa that uses the soap opera *Scandal!* to provide financial education on managing household debt, the evaluators conducted a survey after the financial capability storyline aired. The purpose of the survey was twofold: (1) to determine whether respondents had watched the show and (2) to measure financial knowledge and attitudes. These measurements would then be an integral part of the impact evaluation of the program. The evaluators posted a public announcement in their search for a survey company. After reviewing the various bids, the evaluators settled on a company that not only had experience in data collection and a good understanding of the South African context, but also was familiar with financial capability issues. One of the main advantages of working with this firm, according to the evaluators, was that the company was able to provide useful and informative feedback on initial versions of the questionnaire and how to adapt the questions to the local context. This feedback was subsequently used to improve the survey instrument.

12.3 BUDGET AND FINANCING

Not surprisingly, evaluations are highly resource-intensive, which is an important reason why many programs are not evaluated. The evaluation budget should include both explicit line items, such as data collection, but also implicit costs, such as the value of staff time for all the members of the evaluation team.

Evaluations are highly resource-intensive—an important reason why many programs are not evaluated.

Costs for a sample of impact evaluations funded by the World Bank show that in most cases impact evaluations account for only a small percentage of overall program budgets—about 0.5–15 percent of program costs. However, the variance of costs is high and there are no easy rules of thumb.

Evaluations for smaller programs may be more localized and easier to manage, but at the same time, they may not be able to realize economies of scale of large evaluations. While absolute costs may be lower, the relative cost of evaluating a small pilot program is generally higher than the relative cost of evaluating a large-scale program. Although available resources such as the Living Standards Measurement Study Manual provide estimations of the cost of collecting data, evaluators should first contact the national statistical agency, since any such estimates depend highly on not just country settings but individual teams and project complexity.

BOX 12.2 THE CHALLENGES OF WORKING WITH HIRED SURVEY COMPANIES: AN EXAMPLE FROM INDONESIA

It is not uncommon for evaluators to hire specialized companies to conduct the survey portion of a study. This is especially useful in settings where the evaluation team may not speak the local language or where the scale of the data collection effort precludes the team doing it by itself. While these collaborations are often fruitful, occasional mishaps do take place.

Cole, Sampson, and Zia (2009) conducted an evaluation of a financial capability program that provided financial education to unbanked households on the benefits of bank accounts. The evaluators hired an outside company to conduct a survey to collect measures of financial literacy and behavior. The survey company hired for the data collection portion was initially also responsible for randomly assigning treatments to participants. Participants were given randomly assigned incentives of $3, $5, or $10 for opening a bank account within two months of participating in the financial literacy intervention. The interviewer would draw one of three colored balls from a bag, and whichever ball the interviewer picked determined the incentive. In theory, one-third of households should each have received an incentive of $3, $5, or $10. However, the researchers found that for four interviewers, many more households received the $10 incentive, possibly because the interviewers tried to be helpful to the households. The researchers had to discard all households visited by these interviewers from their evaluation. For the subsequent data collection, the researchers preassigned the incentive amount so that interviewers had no discretion in distributing the incentives.

12.4 ESTIMATING TOTAL EVALUATION COSTS

Because costs can vary so widely, it is best to develop a preliminary budget based on rough but relevant information rather than to proceed on assumptions that may be very far from reality. Important cost categories to consider in estimating total evaluation costs are:

- Staff time (internal and external evaluation staff (advisors/consultants)
- Travel and subsistence costs
- Data collection costs, which can be significant in low- and middle-income countries (LMICs). These costs include the costs of creating and programming (if appropriate) instruments, equipment, training and wages for field staff, their travel (including hotels, and per diem), vehicles/fuel needs, and data entry costs. It is important to carefully consider the assumptions being made to avoid underbudgeting. For instance, in rural settings where access is limited, specialized transportation may be required to transport enumerators; alternatively, in urban settings, additional security measures may have to be put into place.

An example of a budget that outlines these main cost categories is presented in table 12.3. This budget was a part of the evaluation plan of the experimental movie screenings that were used as the main intervention in the Nigerian-based RTF program looking at the role of entertainment in promoting financial education. The budget presented accounts for staff time, travel, data collection (including survey development and fielding), and miscellaneous costs associated with the intervention—such as equipment rentals, venue cost, and incentives like free meals—required for the screenings to take place.

TABLE 12.3 EXAMPLE BUDGET FROM THE NIGERIAN RTF PILOT

ACTIVITY	BUDGET
Field coordinator based full time in Lagos	$30,000
IE team (field supervision and report writing)	$50,000
Listing survey	$15,000
Main survey	$120,000
Travel	$15,000
Movie screenings and incentives	$20,000
Total	$250,000

12.5 EXPLORING FUNDING SOURCES

Although evaluations are typically funded from within a program, from other government resources, from a research grant, or from an outside donor, it is notable that many successful evaluations (including those funded by the Financial Education Fund [FEF] and the RTF) involve resources **beyond** those provided for by a program itself. Given the growing recognition that robust evidence is informative beyond any specific program, evaluation is increasingly supported by existing and new stakeholders with an interest in developing the global base of knowledge on financial capability. Such stakeholders may include governments, development agencies, foundations, and international initiatives, such as the International Initiative for Impact Evaluation (3ie).

12.6 CONTINGENCY PLANNING

An important part of evaluation planning is to carefully consider the risks and challenges that are likely to occur. Many of these happenings are likely to be out of the evaluators' control. However, articulating these risks and a plan for management is an important precautionary step.

It is important to explicitly consider how any possibly deliberate changes in program operations will be handled while the evaluation is under way, especially in the case of newer programs where design may still be fluid.

A simple rule of contingency planning is to build-in a sufficient cushion of budget and time when feasible.

Sound contingency planning should allow for some changes in operations without derailing the overall evaluation, particularly in large multifaceted programs. For example, if a component of a particular financial capability program is modified, there should be procedures for documenting when and how the change occurred and tracking the exposure of particular participants prior to or after the change. A sound monitoring system can be invaluable in this case, and having a solid results framework can then help in interpreting the final results. Indeed, in the best of all worlds, such incidents themselves can be evaluated, thereby shedding light on other program components.

A simple rule of contingency planning is to build-in a sufficient cushion of budget and time whenever feasible. For instance, a common evaluation design is to assign individuals by lottery based on the assumption that a program will be oversubscribed. However, if interest in the program is less than expected, or marketing is less than effective, there may be no oversubscription and hence no basis for the randomization. Even with a purely randomized assignment system, it is always a risk that program participation in new pilots may be so low as to prevent a large enough

sample for analysis. If the evaluation plan is not sufficiently flexible, this could lead to the end of the evaluation.

Then again, with sufficiently flexibility, such problems can be dealt with. For example, in one RTF pilot program for financial education, the take-up rate was extremely low, because of incorrectly targeted marketing. In another instance, the training of educators was inadequate, leading to lack of demand by consumers. However, because of careful oversight of the evaluation and sufficient flexibility to perform a course correction, in both instances it was possible to repeat the marketing and train-the-trainer activities while remaining on schedule and within budget and to do so in a way that maintained the integrity of the evaluation.

12.7 PREPARING AND REVIEWING A FORMALIZED EVALUATION PLAN

A written document that specifies the evaluation design and protocols is a helpful planning, communications, and commitment tool. A clear formal plan is often required by external stakeholders to gain acceptance and permissions, set mutual expectations, and almost certainly as a prerequisite to secure funding. If external evaluators are involved, developing the evaluation plan may be one of the evaluator's chief initial responsibilities and goes hand-in-hand with the ToRs. Even if evaluation teams are purely internal, an overall plan can be critical for managing what might be a program that lasts several years with multiple staff members and reporting requirements.

Here, we provide a checklist (table 12.4) that summarizes key building blocks of an implementation plan.

Evaluation plans should—ideally—be prepared several months in advance and be frequently reviewed. An important part of such reviews in many settings is related to the ethical treatment of human subjects. In some countries and institutions, research conducted on any human subjects requires the approval of a special-purpose board or committee, a process that can be protracted in nature. Relevant instances where such review may be needed include:

Reviewing the ethical treatment of human subjects is a key part of evaluation reviews.

- Research funded by U.S. federal agencies taking place in LMICs must comply with ethical principles set forth in U.S. federal law.
- Academics conducting evaluations may be subject to the ethics review boards at their universities.
- Interventions involving interactions between financial decisions and health that may require ethical review through clinical trials by the local Ministry of Health.

TABLE 12.4 EVALUATION PLAN CHECKLIST

Motivation and objectives of evaluation	X
Major questions to be addressed	X
Program description and results framework model (discussed in chapters 2 and 3)	X
For each objective: ▪ Types of information needed ▪ Sources of information ▪ How sources will be selected ▪ Methods for collecting information (instruments and procedures) ▪ Time frame for collecting information ▪ Methods for analyzing information	X
Evaluation design and methods ▪ For process evaluations, analyses will be primarily descriptive and may involve tabulating frequencies (of services and participant characteristics) and classifying narrative information into meaningful categories, such as types of barriers encountered, strategies for overcoming barriers, and types of facilitating factors. ▪ In the case of impact evaluations, a plan for evaluating participant outcome objectives must include a description of the evaluation design, including a description of the comparison or control group. The evaluation plan will need to specify strategies for encouraging non–treatment group members to take part in the evaluation.	X
Plans for pilot-testing and revising information collection instruments	X
Plans for pilot-testing and revising information collection instruments	X
A comprehensive data analysis plan. The analyses must be structured to answer questions about whether change occurred and whether these changes can be attributed to the program	X
Team practices and procedures for management	X
Timing and scheduling	X
Budget and funding, if available	X
Risks, contingencies, and mitigation plans	X

We note that, regardless of whether ethical review is formally required, the plan should be reviewed with respect to ethical challenges relevant to financial capability program evaluation, which are explored in detail in chapter 13.

Apart from ethical review requirements, external review by a selected group of individuals is desirable and should include the following stakeholders:

- Program or organizational leadership who can determine whether the evaluation plan is consistent with the agency's resources and evaluation objectives
- Program staff who can provide feedback on whether the evaluation will involve an excessive burden for them and whether it is appropriate for program participants

- Advisory board members who can assess whether the evaluation will provide the type of information most important to know
- Participants and community members who can determine if the evaluation instruments and procedures are culturally sensitive and appropriate.

KEY POINTS

Evaluations are the key drivers in ensuring that we are able to understand what went right and wrong in implementing a financial capability program, and whether the program achieved the impacts that it set out to accomplish.

But such evaluations have lots of moving parts and activities that must be done while the program is being implemented. Making sure you have an evaluation plan is critical.

Having a sound plan that carefully considers all the things that are critical to an evaluation and all the things that could go wrong (along with mitigation plans to deal with them) will ensure that evaluation ends up being effective.

FURTHER READING

General

Cole, S., T. Sampson, and B. Zia. 2009. "Valuing Financial Literacy Training." As of October 19, 2012: http://siteresources.worldbank.org/INTFR/Resources/Ziaetl030309.pdf.

Gertler, P. J., S. Martinez, P. Premand, L. B. Rawlings, and C. M. J. Vermeersch. 2011. *Impact Evaluation in Practice*, Washington, DC: World Bank.

World Bank and Inter-American Development Bank. 2010. "Challenges in Monitoring and Evaluating: An Opportunity to Institutionalize," Fifth Conference of the Latin America and the Caribbean Monitoring and Evaluation (M&E) Network, Washington, DC.As of February 26, 2013: http://siteresources.worldbank.org/INTLACREGTOPPOVANA/Resources/840442-1255045653465/Challenges_in_M&E_Book.pdf.

Technical

Herrell, J. M., and R. B. Straw, eds. 2002. *Conducting Multiple Site Evaluations in Real-World Settings*, San Francisco: Jossey-Bass.

Rugh, J., Bamberger, M., and Mabry, L. 2011. *Real World evaluation: Working Under Budget, Time, Data, and Political Constraints*, Thousand Oaks, CA: Sage.

CHAPTER 13

Ethical considerations

When quantitative and qualitative research is done with human subjects, a number of ethical issues may arise when it comes to designing the evaluation and collecting and analyzing data. These issues occur primarily because of the concern that participants who take part in certain types of research may be negatively affected by it—either in terms of their health or social and economic well-being. For these reasons, evaluations must have a robust ethical framework in place for dealing with these risks and for ensuring the best possible outcome for both the research and its subject population.

Above and beyond such overarching ethical concerns, research in developing countries often presents more specific ethical challenges. Vulnerable and underserved populations—such as the poor, children, the elderly, those with little education, and those with mental or physical illness or disabilities—may be or feel coerced, manipulated, or deceived into participating in a research project in developing countries, where oversight and protection systems may be limited or nonexistent. This may also be true in developed countries, of course, but those in developing countries may be at particular risk. These populations may also be more likely to experience negative repercussions from participating in certain types of research and evaluation, including risks to their personal safety, social ostracism, or exclusion from a particular program.

Vulnerable populations in developing countries may be more likely to experience negative repercussions from participating in certain types of research and evaluations.

This chapter discusses these ethical concerns, starting with a discussion of ethical concerns within the context of financial capability programs in particular, and then moving on to more general concerns, using illustrations when possible to illustrate the issues.

13.1 ETHICAL ISSUES IN A FINANCIAL CAPABILITY PROGRAM SETTING

Any program evaluation has to concern itself with issues surrounding human subjects and any evaluation done in developing countries has the additional concerns mentioned above. But evaluations of financial capability programs have

some special ethical concerns, in particular in low- and middle-income countries (LMICs).

First, where education levels and exposure to formal finance are both low, many people may not be able to actually make informed judgments about their own finances. Participants in these settings are likely to be less informed about finances and the implications of their decisions than those administering and evaluating the program—producing a sort of information asymmetry that makes participants more susceptible to deliberate or accidental misinformation.

Where education levels and exposure to formal finance are low, many people may be unable to make informed judgments about their own finances.

For instance, consider a microcredit provider that wants to improve access to credit in rural areas by launching a new marketing program to explicitly promote loan take-up. In such cases, it is important to ensure that field staff are carefully instructed and monitored so they do not inadvertently promote irresponsible borrowing, especially if households are not familiar with the terms and penalties associated with repayment.

Second, financial capability programs may introduce participants to new risks, which can affect their well-being and may have long-term effects on financial decision making and other aspects of household behavior. Of course, not all risks are to be avoided. But it is important to recognize and consider the possibility of negative consequences and to weigh them accordingly.

For instance, in the case of the microcredit provider above, it is important to note that when they promote entrepreneurship, they inherently expose households to the risk of starting a new business, which may or may not be appropriate for certain households.

Another source of risk may come from the fact that some countries' financial and legal structures are underdeveloped, which could leave participants with limited access to insurance, consumer protection laws, or other methods of formal redress. Such underdeveloped structures make participants more vulnerable to any adverse situations that might emerge from the program or evaluation.

Underdeveloped financial and legal infrastructures may leave participants with limited access to insurance, consumer protection laws, or other methods of formal redress.

Consider, for example, a training program in which participants receive general advice on how to save their money in a formal financial institution. Such advice seems innocuous enough, but what if the banking environment is largely unreliable and unregulated? If that's the case, then participants may be exposed to the risk of fraud or loss.

Alternatively, if unscrupulous parties obtain and use participants' financial data for unlawful gains, households may not have the legal means to address the situation. In some more extreme cases, these risks can even be physical. For instance, in post-conflict settings where the rule of law is still being established, communi-

ty-based banking schemes in remote areas may become a target for crime, corruption, and/or coercion unless special care is taken to secure their administration.

It is also important to consider risks that may be more marked for certain subgroups of participants. For example, in some societies, introducing financial capability programs for women may improve their bargaining power and welfare, but could also lead to risks, such as social and even physical harm.

Given these ethical concerns, evaluators should be mindful of them when undertaking a study and seek to mitigate their effects by building safeguards into their research and evaluation design; such safeguards include strict procedures for recruitment, informed consent, and confidentiality procedures, among others.

Thus, a first-order concern when performing any financial capability program evaluation should be to protect participants' integrity, privacy, safety, and human rights through a number of safeguards. Below we discuss a number of these safeguards and considerations in greater detail.

13.2 INFORMED CONSENT

One of the most basic safeguards in research involving human subjects is what is referred to as **informed consent**. Potential participants in your evaluation (whether they receive the program itself or are part of the control group) should be given the opportunity to give their consent to participate, based on fully available and easily accessible information. Importantly, they should also be able to decline the opportunity to participate in the research (or to withdraw at any point during the study) without the fear or risk of adverse consequences for them or their families. In an evaluation, this means that participants should be able to refuse to be interviewed or to take part in a survey without risking being dropped from the program altogether.

Informed consent is one of the most basic safeguards in research involving human subjects.

What constitutes informed consent? At a minimum, participants need to know the following:

- What the purpose of the research is
- Who the research is for
- Who is conducting the research and how to contact them or their representatives if necessary
- How the information collected will be used
- Who will have access to their personal information
- What they will be asked to do or discuss

- How much of their time will be required for participation in the evaluation (e.g., to complete surveys or attend focus groups) and over what period of time
- Where the research will take place (the person's home, a local school or hall, another village, a particular area of the city, etc.)
- What risks and/or benefits (including compensation for their time) are involved through their participation in the evaluation.

Of course, there are other things that might be asked as part of the informed consent process, but the above list constitutes the bare minimum that must be part of it.

When seeking informed consent, evaluators should provide the necessary information to participants in a format that is complete yet also understandable and meaningful to them and within time frames that suit both the participant and the study. It is not useful to provide information, even if complete and accurate, if it is written (or communicated) in a way that participants do not understand or that does not make sense to them. Nor is it useful to do this so quickly that participants cannot process what they are being given or asked to do.

The most common way of documenting consent is to have participants (or their legal guardians, if appropriate) sign a "consent" form by which they agree to take part in the evaluation. It is also customary for minors to also sign "assent" forms, which are nonbinding but show that the youth is knowingly participating.

Obtaining **active, written consent** is the most transparent way of holding an evaluation accountable. But an alternative is **passive, verbal consent**. In this format, research administrators would develop a description of the research effort that includes all the elements from the above bullets and would read it to potential participants exactly as written so that there is no variation depending on which member(s) of the research staff may be recruiting the participant. Once research administrators have finished reading the script, participants may then ask any questions they might have and are given the opportunity to verbally refuse participation, without a consent form. In some cases, verbal consent is more appropriate for literacy-related reasons: Those who have low literacy may not be able to read or fully understand the consent form, nor may they be able to meaningfully provide written consent.

In dealing with people with low literacy, even basic financial terms may not be understood and should be defined in the consent process.

Whether informed consent is written or oral, it is critical to realize that in dealing with people with low literacy, even basic financial terms—such as "interest" or "budget" may not be understood and should be defined in the consent process. It is well-documented that access to financial services in the developing world is extremely low, so basic financial terms are likely to be unknown to many participants.

Another important consideration is that, in certain circumstances, getting informed consent may be difficult or have adverse consequences. For example, school-based financial capability interventions that are dealing with adolescents may be unable to proceed if evaluators can't get informed consent from the students' parents, or if the evaluation or intervention covers financial or other matters that adolescents want to keep hidden from their parents (such as how they spend their money). In some contexts, financial matters may be considered so sensitive that obtaining informed consent for a program and evaluation may be difficult, because signed informed consent may feel intimidating or threatening to the subject. Finally, lengthy informed-consent procedures that are too burdensome may also deter participants.

While it is important to bear these risks in mind when designing research with human subjects, the principle of informed consent may, and often does, yield positive outcomes. In program evaluations, for example, seeking informed consent from subjects may improve the quality of data obtained because subjects are more willing and prepared to provide the required information. It may also increase participation rates, because subjects may be more inclined to participate in activities they understand well and for which they receive credible assurances. It can also be a helpful process in and of itself, because it allows evaluators to reflect critically on their relationship with their subjects and the implications to the subjects of participation in the study.

13.3 CONFIDENTIALITY

Ensuring confidentiality is key to developing trust and to getting the most useful information from participants. In designing your evaluation, one key issue to consider is how subjects' personal information will be used, and by whom; whether there is any type of information about which confidentiality cannot be guaranteed (such as if the evaluation uncovers evidence of illegal activities); how the data will be stored; who will have access to it; and how long it will be maintained.

Ensuring confidentiality is key to developing trust and getting the most useful information from participants.

Data safeguarding is an important issue in any research with human subjects, especially in financial capability programs when personally identifiable information related to financial history or transactions is collected. This is the case even if you expect to "de-identify" the data for analysis and dissemination—that is, strip out any information that could identify the individuals involved. A system for securely storing and restricting access to personal or proprietary data should be developed, as well as a plan for storage, use, and destruction of the data upon completion of the evaluation.

Table 13.1 lists some common questions that should be part of any data-safeguarding plan.

TABLE 13.1 ELEMENTS OF DATA-SAFEGUARDING PLANS

AREA	ELEMENTS
Data sensitivity	▪ Are data de-identified (i.e., no personal identifiers that allow users to determine the individual's identity)?
Responsibility for data safeguarding	▪ Who has overall responsibility for data safeguarding? ▪ Who else will have access to the data? ▪ Will all who have access to the data be trained in appropriate safeguarding procedures?
Data safeguarding procedures	▪ Who is responsible for recruiting/enrolling participants? ▪ Are unique identifiers assigned? ▪ What (if any) personal details will be recorded? ▪ Where are data recorded? ▪ Will there be copies of the data (hard copies, soft copies, web-based storage, etc.)?
Data transmittal	▪ How will data be transmitted, for example from program sites to evaluation "headquarters"?
Data disposal	▪ How and when will data be disposed of when the evaluation is complete?

A common question in program evaluations relates to who owns the personal data collected by the program. Do the data "belong" to the researchers, to the funders of the evaluation, to the funders of the program being evaluated, to the implementing agency, or to all the above? It is important to consider this question explicitly, especially given that who owns the data may have important implications for participants' privacy and willingness to participate in the study. For instance, suppose an employer works with a commercial bank to run a financial education program in their workplace. For evaluation purposes, extensive financial information on participants may need to be collected. Participants may be more willing to provide private information on this to a third-party evaluator than to their employer, but might not want it shared with a financial institution that could exploit the information for unwanted marketing or other purposes.

13.4 ANONYMITY

Full anonymity may offer the best protection of participants' privacy and guarantee more accurate information, but it may not be feasible or desirable.

Yet another important question to consider is whether any personal information (such as names, locations, and other data that could identify specific participants) will be changed or hidden. While full anonymity may offer the best protection of the participant's privacy and guarantee more accurate information, it may not be feasible or desirable.

Anonymity may be difficult to guarantee, for instance, when evaluations are taking place in small villages or communities where people's activities (such as participating in a survey) and basic information are well known to neighbors. Moreover, in some cases, information collected about participants cannot be de-identified because of

the needs of administrative databases. The key in these contexts is to be clear with participants about how much anonymity is guaranteed by specific protocols when consent is being sought and to strictly maintain these protocols.

13.5 RISK ASSESSMENT AND MITIGATION PLANNING

While some risks in conducting a financial capability evaluation may be pretty obvious from the start, some may not be nearly as obvious. And while some risks may be unavoidable, evaluators should make a concerted effort in designing and implementing a program to achieve the fullest possible understanding of the risks that a program may pose for both researchers and participants. Risk assessments should consider all the ways an evaluation may harm participants and others (researchers themselves, subjects' families, etc.) and, if the evaluation involves a new and untested intervention, the potential risks from the intervention itself.

Evaluators should make a concerted effort to achieve the fullest possible understanding of the risks a program may pose for researchers and participants.

Important issues to consider should include financial and economic harm as well as other types of harm.

- **Could the research compromise the physical safety or lead to psychological stress for researchers or subjects?** In social science research, this is relevant where the research topic is sensitive (such as human trafficking, domestic violence, gang activity, etc.) or when the research would be conducted in areas with violent conflict or high levels of crime. This is not likely to be relevant for financial capability programs, but evaluators should at least consider the potential for harm.
- **Can participation in a program evaluation negatively affect the subjects' continued participation in the program?** This might be the case when program administrators and/or in-country staff incorrectly assume that participation in an evaluation precludes participants from benefiting from the program once the evaluation is completed.
- **Can participation in a program evaluation have social repercussions for subjects?** Examples of this would be ostracism by members of the community or program staff and exclusion from other programs or services.
- **Can participation in the evaluation negatively affect the participants' well-being?** Consider the example laid out earlier in this chapter of a financial capability program that improves access to credit. Taking on credit-based debt can open new opportunities for households, but it also exposes them to financial risks.

While assigning some people and not others to an intervention has ethical implications, randomization can sometimes be more "ethical" than it may first appear.

As noted above, clearly communicating any risks that participation may entail is an important part of obtaining true informed consent. Evaluators are also responsible for minimizing any risks to themselves and their subjects—a responsibility that includes measures to arrest or address any problems that may emerge. For instance, providing information about financial distress to evaluators can raise the risk of social harm or stigma. The evaluator is responsible for maintaining data protocols for confidentiality and protection, and for mitigating the consequences of any data breach.

13.6 ETHICS IN EVALUATIONS WITH RANDOMIZED INTERVENTIONS

Randomized interventions often present a common ethical conundrum: program assignment. One concern is that it may be "unethical" or "unfair" to randomly assign certain people to a purportedly helpful program while excluding other equally needy candidates or, conversely, to assign the latter group to something unknown and possibly harmful.

For example, giving people in LMICs credit can be positive or negative, depending on their ability to repay, the likelihood that they will get other (informal) credit to repay it, and other factors. The conundrum in these cases is the ethical implications of giving this intervention to some people but not others.

While this is an intuitively understandable ethical concern, randomization can sometimes be more "ethical" than it may first appear. First, an evaluation involving randomization allows evaluators to ascertain whether a certain financial capability program is indeed beneficial, a situation that is an improvement over rolling out a program to all eligible participants based on assumptions and anecdotal evidence. Given scarce resources, randomization is warranted: We first randomly allocate a benefit, rigorously assess its impact, and then determine whether it is worth rolling out more widely. If the evaluation reveals that a program is not having the intended positive impact, its resources can be used to redesign the program or for a different kind of intervention.

For example, a randomized evaluation of a financial capability workshop may reveal that the workshop materials or structure are not achieving the intended outcomes (such as improving participants' ability to budget their income). Rather than having rolled it out widely at much greater expense, and wasting many more participants' time in an ineffective program, a randomized evaluation enables those responsible for the program to modify the intervention and increase its impact.

BOX 13.1 DEALING WITH DISAPPOINTMENT IN CONTROL GROUPS

The Russia Financial Literacy and Education Trust Fund (RTF) pilot program in Brazil, which consisted of a school-based financial capability intervention, offers a good illustration of some of the ethical challenges of randomization. The design of the program was to invite all schools that were interested in the financial capability intervention; from within those, the evaluators chose the treatment and control groups. As a result, many of the control-group schools that were interested in the program did not get it, and a number of these were upset because they did get the intervention. The evaluators carefully explained to them that they would get it in a couple of years and that the program they got then would be even better because of the lessons learned from the randomized evaluation.

Second, in some settings, randomization can, in fact, be the fairest method of assignment. For example, in situations of limited resources or when the roll-out will be phased, randomization can ensure all eligible people are equally likely to receive a benefit for which there is excess demand. In the case of phased roll-out, randomizing which groups or communities will receive the benefit first takes advantage of the phased roll-out by creating a robust evaluation design.

Making sure the control group is not made worse off than they were before the study is a key concern in randomization: The treatment group is offered a "benefit" while the control group is offered either a lesser benefit or nothing—neither of which should make them worse off than they were before.

It is also critical to put in place mechanisms and safeguards to prevent conflict (e.g., within a community) that arise from randomization. One way of doing this is to hold a public meeting where the evaluation is clearly explained and people can ask questions about the aims, design, and implications of the research. When a more involved intervention is needed to deflect conflict, it may be possible to conduct the random selection of participants in public so people know randomization is legitimate or to extend the treatment condition to the rest of the community at the end of the evaluation.

It is critical to put in place mechanisms and safeguards to prevent conflict that may arise from randomization.

At any rate, the safeguards in place should be appropriate for both the community and the requirements of the evaluation. The public announcement and commitment to a future schedule of implementation, with the support of local community leaders, is an important step. Moreover, such an announcement can be good for the design because it can help prevent contamination by helping ensure that control groups are not exposed to the treatment condition or to aspects of it.

BOX 13.2 RTF IN PRACTICE: EXPLOITING PHASED ROLL-OUTS IN NIGERIA

Two RTF projects in Nigeria take advantage of programs that are national in scope but staged to roll out in phases. The first project consisted of the development of a feature film dealing with financial capability issues—a part–social marketing, part–education entertainment initiative. The film is scheduled to air first in some parts of the country, allowing evaluators to use control groups for evaluation, where the control groups are those areas where the film has not been shown yet. This approach ensures that the film is eventually shown across the country for the benefit of any and all who might watch it, while also allowing for an experimental design.

Another project investigates the impact of the "I-Save I-Win" (ISIW) promotion, a large nationwide financial capability program launched by Nigeria's InterContinental Bank in the spring of 2011 with the aim of mobilizing precautionary savings through mass market savings accounts. The ISIW promotion is a heavily publicized lottery incentive scheme designed to educate consumers about savings through first-hand sustained use of individual savings accounts. By exploiting the staggered introduction of various components of the ISIW promotional campaign over time, the study will measure the marginal increase in savings applications and banking activity following each component of the media campaign.

13.7 OTHER ETHICAL OBLIGATIONS

In addition to considering the welfare of research subjects, there are broader ethical obligations, including objectivity, transparency, integrity, fairness, and professional competence. They can ensure that evaluation findings are reliable, trustworthy, and empowering to stakeholders who can learn and follow up, to donors who seek accountability, and to the broader public.

Objectivity, transparency, integrity, fairness, and professional competence can help ensure evaluation findings are reliable, trustworthy, and empowering.

In this context, it is important to be aware of potential **conflicts of interest** that could arise between the parties involved—such as funding agencies, implementers, researchers, and evaluators—and the research subjects. An example might be a financial education program that promotes a product or service that the funding agency has a financial stake in. The issue of objectivity can be a particular challenge, especially if evaluators either belong to, or are contracted by, the implementing and/or funding agency. If internal experts are performing the evaluation, it is important to foresee and prevent ethical breaches resulting from attempts by parties involved in implementing or funding to try to control the content and dissemination of findings or change the terms of the evaluation mid-stream. In such circumstances, maintaining objectivity and independence can be difficult, especially if emerging results are not as expected or hoped for.

Setting out basic **rules of engagement** at the outset of an evaluation or working with an external evaluator can prevent problems down the line. These could include, for example:

- Specifying roles of funders and the implementing agency in the evaluation
- Establishing clear lines of communication between evaluators, implementers, funders, and other relevant stakeholders
- Articulating and committing to the evaluation's standards for transparency and accountability
- Describing any known or potential conflicts of interest.

Researchers have an ethical obligation to form partnerships in the countries in which they work to foster the development of a research infrastructure and contribute to capacity building.

Researchers have long recognized that resource constraints—both human and financial—limit social research and program evaluation in developing countries, seriously hampering these countries' ability to build research and evaluation capacity. As such, one could argue that researchers have an ethical obligation to form partnerships in the countries in which they work—an activity that can foster the development of a research infrastructure and contribute to local capacity building.

At the heart of this contention is the idea that countries have the right to benefit from hosting research. In social research and program evaluations, countries should benefit directly from the results of studies that can help inform policy or improve initiatives that reach out to underserved populations. As such, researchers should leverage the opportunity for research and evaluations to build local capacity in these areas as part of the wider ethical obligations of researchers to their host societies. In the medical sciences, this is articulated explicitly in the Helsinki Declaration and the Council for International Organizations of Medical Sciences Guidelines, which identify the contribution of research to capacity building as a key benefit of research to host countries.

Activities geared toward capacity building could include efforts to strengthen research capacity, to bolster the institutional framework needed to support local research efforts, and to develop robust ethical guidelines. These can involve partnering with local researchers and providing training or advice to local institutions.

KEY POINTS

Ethical considerations should be fundamental to doing evaluations, both to ensure that human subjects and the data generated from them are protected and to ensure that the evaluations themselves are as successful and useful as possible. They should not be seen as marginal to the evaluation or as obstructions to efficient research activities. Evaluators should reflect on the ethical implications of their studies early on, and continue to monitor them throughout the life of the project.

Some countries have adopted various approaches to or standards of ethical conduct for research among their population. Examples of this include the Social Sciences

and Humanities Research of Canada and the National Committee for Research Ethics in Norway. Evaluators should be aware of these standards or approaches—where they exist—so they can be followed and/or incorporated into the ethical frameworks of their own studies.

Still, if research will be conducted in developing countries or countries that lack such standards and oversight, researchers should assess whether anything they will be doing there involving human subjects would be possible in developed countries. If the answer is "no," they should think carefully about the ethics of proceeding with the evaluation in its current form. The absence of formal requirements and rules about ethical professional conduct in social research and evaluations should not preclude an evaluator's duty of care to her subjects.

As the use and evaluation of financial capability programs becomes more widespread, it is useful not only to keep abreast of developments in the field (which can help you think of the ethical challenges you may face in your own evaluations) but also to disseminate the experiences and results of your own study, which can help build a robust evidence base built on ethically sound research.

FURTHER READING

General

USAID (U.S. Agency for International Development). 2008. "Procedures for Protection of Human Subjects in Research Supported by USAID." As of February 26, 2013: http://www.usaid.gov/policy/ads/200/humansub.pdf.

UNEG (United Nations Evaluation Group). 2007. "Ethical Guidelines for Evaluation." As of February 13, 2012: http://www.unevaluation.org/ethicalguidelines.

Technical

Angell, M. 2000. "Investigators' Responsibilities for Human Subjects in Developing Countries," *New England Journal of Medicine* 342 (13): 967–69.

National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. 1988. *Belmont Report: Ethical Principles and Guidelines for the Protection of Human Subjects of Research*, Washington, DC: Government Printing Office.

Varmus, H., and Satcher, D. 1997. "Ethical Complexities of Conducting Research in Developing Countries," *New England Journal of Medicine* 337 (14): 1003–05.

CHAPTER 14

Documenting and communicating results

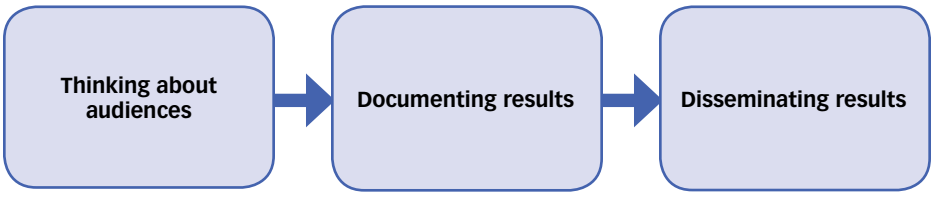
This Toolkit has laid out how monitoring and evaluation takes place for financial capability programs—on how to set up a monitoring system and the different types of evaluations—process, impact, and cost—in order to yield results that are useful in telling program staff, funders, and other stakeholders how effective or ineffective a program is in meeting the aims set out for it.

But effectively conducting monitoring and evaluation is not the end of the process. Equally important is making sure that the program effort gets documented and that the results are communicated to those who need or want to see them. The results of evaluation efforts, regardless of how compelling they are, will not be useful if no one sees them. Moreover, even if the results are seen, they will not be as useful as they could be if they are not appropriately communicated to the specific audiences that need them. In other words, the results need to be understandable.

The results of evaluation efforts, regardless of how compelling they are, will not be useful if no one sees them.

This chapter discusses how to document and communicate financial capability evaluation results effectively. After discussing the importance of documentation, it presents key considerations in thinking about how best to write up your findings, how to think about the key audiences you want to inform, and how to communicate with these audiences effectively. The discussion is organized around the flow chart shown in figure 14.1.

FIGURE 14.1 DOCUMENTATION AND COMMUNICATIONS FLOW CHART



14.1 THE IMPORTANCE OF DOCUMENTATION

An early innovator in program evaluations argued that evaluation is fundamentally about learning. Ideally, this learning will advance knowledge in the field and inform a range of practical decisions, both inside and outside of your organization. Of course, foremost among the lessons that emerge from a financial capability program evaluation is knowledge about how well your program is working. The best way to ensure that such learning results from your evaluation is to document and communicate your evaluation's results. Documenting and communicating your results is the key to ensuring that interested audiences—including program participants and managers, members of your organization, funders, and organizations that operate similar programs—can learn from what you found in your evaluation and apply the results.

As discussed in chapter 1, the base of evidence on what makes for an effective financial capability program is very weak. Despite the large numbers of financial capability programs and interventions, few have been evaluated at all and, where they have, the design of the evaluations has tended to limit what others can learn from them.

Beyond whether or not programs have been evaluated is the more general concern that even when they have been, evaluations were not always documented. Failure to document and communicate the results of an evaluation obviously impedes the ability of others to learn from it. But there are even more insidious results—such a failure to document can contribute to a significant problem in evaluation science: the problem of "publication bias."

The evidence base should include all types of results, not just those that found dramatic effects, so that a balanced and accurate picture of program or intervention effects will exist.

Publication bias is the tendency of researchers and evaluators to publish results that only show large intervention effects, also called the "file drawer effect." The term is based on the assumption that many studies that have found weak effects, no effects, or negative effects remain unpublished—that is, stuck in a researcher's file drawer. If part of the purpose of a financial capability program evaluation is to contribute to the evidence base for a particular kind of program or intervention, it is important that that evidence base include all types of results, not just those that found dramatic effects, so that a balanced and accurate picture of program or intervention effects will exist. The best way to ensure that key audiences learn from your evaluation is to avoid the "file drawer effect" by documenting and publishing the evaluation results, no matter what they say.

14.2 THINKING ABOUT THE RELEVANT AUDIENCES

Going back to figure 14.1, the first step in documentation is thinking about or considering the relevant audiences for that documentation. The tendency for many is to

jump right in and begin documenting, but the end result of doing so may be the creation of documentation records that do not suit the needs or interests of those who will be reading them.

Documentation needs to consider the intended audience(s) and the communicator's purpose in communicating with that audience(s).[1] There are likely to be many audiences, and each may have very different needs for the documentation. One way to identify the audiences for the evaluation results is to answer a series of questions:

- What audiences will be interested in your evaluation results?
- What information will these audiences look for?
- How will they use this information?

Documentation must consider the intended audience(s) and the communicator's purpose in communicating to that audience(s).

One way to begin answering these questions is to think about your audiences in terms of whether they are internal to your organization or external.

14.2.1 Internal audiences

Program staff and **participants** are perhaps the most obvious internal audience that will be interested in the evaluation findings, especially those in management roles. People who operate the program or participate in its day-to-day activities can use the evaluation results to understand how well the program is performing. Program participants of course have an interest in understanding what the program effects may be, especially for outcomes that are not immediately visible to them such as long-term changes in on economic status.

If the evaluation identified problems, program staff are the ones who will likely need to fix them.

More specifically, for program staff and participants who played a role in the evaluation, they may have provided input to documenting the evaluation's results and will want to know that their input was captured accurately. If the evaluation identified problems, these are the people who most likely will need to fix them. In addition, if the evaluation was critical of the program or some aspect of it, program staff will want to know that the discussion of problems was handled constructively and with tact and sensitivity. If the program is particularly successful, program staff may be called on to help scale the program up or implement it in some other settings. If the evaluation was formative, they will need to understand the lessons learned and decide whether and how to modify the program to incorporate those lessons. They may also take an interest in the evaluation's approach for reassurance that the analysis was rigorous and objective.

In the case of financial capability programs, needed changes could involve modifying intervention materials or even delivery mechanisms. They may also involve making

[1] Adapted from James Kinneavy, *A Theory of Discourse*, New York and London: Norton Publishing Co., 1971.

staffing changes, such as training additional staff to deliver a particular service, or retraining staff to improve their effectiveness in delivering a particular service.

Not surprisingly, given the hands-on roles such program staff have, they are going to be especially interested in a more detailed discussion of the program.

Upper management is likely to focus on the "short story" of program outcomes and key takeaway messages.

It is important to consider that **upper management in the program** may face decisions that are different from those facing program staff. Therefore, upper management may have different needs in what they want from documentation. Managers may be charged with deciding whether to continue supporting and investing in the program or to end it; or in the case of a successful program, they may be charged with deciding whether to expand it or replicate it in some other setting.

Confronted with these decisions, upper management is likely to focus on the "short story" about program outcomes and key takeaway messages; correspondingly, they will be less interested in process issues surrounding the program or methodological details about how the evaluation reached its results. In particular, they will be interested in what they are being asked to do—that is, the action items that the evaluation identified and how these emerge from the evaluation results.

14.2.2 External audiences

There are also key external audiences for evaluation results, chief among them being funders, policy makers, researchers, and people who operate similar programs.

Funders may include the financial sponsor or an organization that contributes more broadly to the program or organization, such as a foundation or other nongovernmental organization (NGO). Like internal senior management, sponsors or other funders face difficult decisions about what kinds of initiatives they want to support and what types of programs or interventions are effective in accomplishing the goals they want to support. They will take an interest in whether the program is performing effectively and thus should be maintained or scaled up, or, conversely, whether it is underperforming and requires modification or should be terminated.

Funders are likely to be interested in bottom-line results but lack the time or inclination to read them in detail.

They are thus likely to be interested in the bottom-line results but lack the time or inclination to read them in detail. Having said that, they may be less familiar about the specific details of the program, which means they may need more context than internal audiences. This suggests that funders will also be interested in a high-level summary that goes light on methodology, sets the context for the evaluation, and emphasizes results. They will also want to know what actions—i.e., recommendations—follow from the evaluation. Finally, they will be interested in information about cost-effectiveness. In sum, funders need the results presented to them in a way that ties the results to proposed action steps that you recommend funders should undertake.

Policy audiences can include a range of stakeholders who are interested in financial capability programs. This set of audiences could be national, regional, or local policy makers who want to understand the effectiveness of programs intended to address financial capability. These audiences can also include community groups, advocacy organizations, or other parties that have an interest in the issues addressed by a program.

Policy audiences will view results through the prism of real-world problems.

Like funders, policy audiences are interested in evaluation results. They are likely to be less interested in program-specific actions or recommendations that emerge from the evaluation. Instead, they are likely to be more interested in how your results are relevant to their interests and the broader implications of your results for similar programs that address similar issues. Like funders—perhaps even more so—policy audiences will view your results through the prism of real-world problems, and they are likely to bring the information they get from your results to bear on a problem they want to address.

Researchers will be interested in what the evaluation adds to the literature or evidence base.

A third potential external audience consists of **people involved with similar programs**. This audience will be interested in the details of your program and will also want to know in some depth what the purpose of the evaluation was, what key questions were asked, what the approach to data gathering and analysis was, and what the results imply for modifying or expanding the program. They are most likely to use the information from the evaluation as it applies to their own programs—to make judgments about the findings' relevance and to draw lessons for operating their programs.

Finally, **researchers and evaluators** are likely to be most interested in what your evaluation can add to the literature or evidence base on the subject at hand. They are likely to focus on the methodology and the validity of the evaluation results, the strength of evidence that inheres in the results, and how the results add to or modify what is known about the type of program under evaluation. They are likely to view your evaluation primarily as it speaks to research issues. They are likely to use the information to add to their knowledge about a given field or discipline, in the evaluation sciences or elsewhere, and perhaps to pick up some guidance about how to conduct a similar evaluation. For such audiences, journal articles can be a logical way to document the research and communicate it.

14.3 COMMUNICATING WITH KEY AUDIENCES

Starting out by thinking about whom the particular audiences for the evaluation will be—and what their communication needs are—is an invaluable guide in determining the best means of communication. Documenting the results of the evaluation is the next step, as shown in figure 14.1 above.

If anything, formal reports are useful for accountability purposes.

The most commonly used way of reaching all these audiences is a **formal report** that includes a 2–4 page summary, which can potentially stand alone as a document. Admittedly, there are differences of opinion about whether a report (or any single form of communication) is the optimal way of reaching multiple audiences. But this debate often presupposes that "report" refers to a thick, dry, technical report that will go unread by many key audiences, especially the busy managers in your own organization and funders or policy makers.

Despite these concerns, there are occasions when a formal report is appropriate. Indeed, in many cases, if only for accountability purposes, it is likely that you will be expected by both upper management in your own organization and program sponsors to prepare a formal, comprehensive report. Given this expectation, it is important to think about how you can use this expected and often mandatory deliverable to effectively address the audience or audiences you want your evaluation results to reach.

In this section, we offer some guidance on how to plan, structure, and draft an effective evaluation report. We suggest that reports do not need to be thick or dry and in fact will be most effective if they are not overly long or technical. In addition, based on your thinking about the audience or audiences for your results in the first step, a report and the accompanying summary can be tailored to meet audience needs. Finally, depending on which audiences you are most interested in reaching, reports can be supplemented by other forms and venues of communication that can enhance the effectiveness of your communication effort.

Table 14.1 contains an outline of what a formal report should contain.

We explore each section in the table in more detail and discuss relating the contents of the section to the needs of the audiences you are addressing. In addition, we discuss how other products or forms of communication can feed (or draw from, depending on the order of composition) from the sections of the report.

14.3.1 Summary

A summary is critical for reaching high-level managers, sponsors, and policy audiences.

The **summary** is a critical document for reaching high-level managers inside your organization, as well as sponsors and policy audiences. Ideally, a good summary will motivate these readers to delve more deeply into your report, but more often than not, the summary is the only place these audiences will consult to learn about your evaluation results. And as noted above, these readers all have a practical orientation—they want to find information that will help them make decisions about continuing, expanding, or ending the program or similar efforts based on your results. So the summary needs to address hard-nosed questions such as the following:

TABLE 14.1 GENERIC TEMPLATE FOR A FORMAL REPORT

SECTION	CONTENTS
Summary	▪ Brief discussion of context, approach, key findings, and recommendations for action
Introduction	▪ Brief overview of the program ▪ Purpose of the program ▪ Brief overview of the approach ▪ Evaluation stakeholders ▪ Evaluators and their relationship to stakeholders ▪ Overview of the contents of the report
Program Description	▪ Program history, background, and development ▪ Program goals and objectives ▪ Program participants and activities
Evaluation Design and Methods	▪ Evaluation questions ▪ Data collection methods used to address each question ▪ Analysis methods for each type of data collected
Findings and Results	▪ Description of how the findings are organized (e.g., by evaluation questions, themes/issues, etc.) ▪ Results of analyses of quantitative or qualitative data collected (usually represented in tables, graphs, or other visual illustrations, and text)
Conclusions and Recommendations	▪ Conclusions drawn about the evaluation results ▪ Recommendations for action based on these conclusions ▪ Suggestions for further study, if applicable

- **Why should I care?** What motivated this program? How is the program addressing an important problem and how does the evaluation shed light on the program's contribution to addressing it?
- **So what?** How is this information going to help me make a decision?
- **What is your point?** What are the key results from the program evaluation?
- **What am I supposed to do with this information?** How do the key results translate into action?

As noted in the table, the goal is to be brief, but it is also important to realize that this might be the only part of the document that busy policy makers read, and therefore it must be sufficient enough to stand alone for those who will read nothing else. That doesn't mean it must be long, but it does mean that it must have enough context for someone to understand what the program was about and why it was conducted. It is tempting to simply start with the findings, but the findings will have little meaning if the audience doesn't understand the context. Having some discussion at a very high level is also useful in this regard (for example, indicating that the program was evaluated using a mix of quantitative and qualitative methods).

A summary should be brief but sufficient enough to stand alone for those who will read nothing else.

14.3.2 Introduction

The **introduction** sets the context for the evaluation. It should describe the program and the problem it addresses, explain why the evaluation was conducted, provide some overview of the approach used, and give the reader an overview of the rest of the report—a sort of roadmap for what's in the rest of the report as a tool to orient the reader. The roadmap part of the introduction can also serve as an opportunity to guide readers in how to read the document. Given that a document like this will serve multiple audiences, it may be helpful to direct those who don't need all the detail to skip some chapters that may not be useful to their purposes. The roadmap can also direct readers to particular appendixes that go into more detail on certain subjects.

An introduction should provide a roadmap for what's in the rest of the report.

If the primary audience consists of a particular stakeholder group, such as the sponsor or a policy audience, it may also be helpful to identify the evaluation stakeholders and describe their relationship to the evaluation (i.e., did they commission it? Are they expecting to use the results to inform a particular decision or set of deliberations?).

14.3.3 Program description

The typical **program description** chapter presents background information on the program's origins, goals, objectives, participants, and activities as developed during the conceptualization phase discussed earlier in the Toolkit. As noted earlier, it is sometimes the case that evaluation reports create the first formal description of a program. Such a description is valuable in its own right as a record of what the program does; in addition, if the evaluation focuses on processes and implementation, the program description becomes especially valuable.

However, if the primary audience is familiar with the program and the only uses of a program description are likely to be archival, this section could be created as a separate appendix or generated informally as an internal record of the program's design and activities.

14.3.4 Evaluation design and methods

The three main pieces of the **evaluation design and methods** are:

- Evaluation questions: What specific questions did your evaluation address?
- Data collection methods used to address each question: How did you collect data to address these questions?
- Analysis methods for each type of data collected: How did you analyze the data you collected?

As noted above, some overarching discussion of this material should appear in the introduction to provide enough context to understand what was done; here, that material would get expanded treatment. If you expect that your primary audience will be other researchers or evaluators, you may wish to give this material more detailed treatment than you would if your main audience consists of policy makers, who will likely not be as interested in the methodological details. Even in the latter case, however, one reason to give careful thought to this section is that you can use your evaluation questions as a way to structure the results that follow in the next section.

14.3.5 Findings/results

The **findings/results** section presents the key findings from the evaluation. "Writing up" the findings from an evaluation is not always as straightforward as it might seem, and it is not the same as conducting analyses that are part of the evaluation. Conducting analyses are, pretty self-evidently, an "analytic" (or "decomposition") activity in the sense that assessments are broken down into small pieces, usually at the task level.

"Writing up" the findings is not always as straightforward as it might seem.

Writing up the results, however, is a "synthetic" activity—one that involves putting all the resulting findings together in a way that makes sense and leads to the conclusions and recommendations that follow. That could be aligned with the specific tasks, but it may not be. As noted earlier, it may be aligned with the evaluation questions, which, in turn, may each be composed of multiple tasks or pieces of tasks.

Particularly if the analysis is quantitative, the writing itself may be a significant piece of the analysis. The writers may need to make some interpretive decisions, particularly if they wish to translate the writing into themes or messages that lend themselves to translation into action.

Adding graphics to display key results can make a findings section more reader-friendly and accessible.

To make this section as reader-friendly and accessible as possible, consider the use of visual illustrations, such as graphs or tables, to display key results. Particularly for quantitative findings, simple charts like bar graphs or pie charts can easily convey quantitative relationships that would require a good deal more space to communicate with words.

However, while graphics are very compelling, they are not going to be very useful if they are poorly designed or overly complicated. Taking the time to consider how to create graphics will go a long way toward making them more accessible and, thus, more useful. This often means removing unneeded grid lines, making the labels and keys clear (as opposed to what was added into the spreadsheet to generate the graphic), and labeling something if it makes sense to do so.

Using qualitative or conceptual graphics can also be very valuable.

Beyond whatever is done to make the graphic accessible, it is just as important to make the text description that goes with that graphic useful to readers in helping them interpret the graphic. Graphics, especially quantitative ones, impose a burden on readers; if the graphic is unclear and/or the description is unclear, they can actually be counterproductive, both because they require effort to process (which can lead to frustration if readers can't process them) and because they take readers away from the linear argument being made in the text. If a reader's attention strays to the graphic, he or she can lose track of the argument in the text.

A final point on graphics is that qualitative or conceptual graphics also make sense, both in this section and in others. For example, flow charts, like those that capture the evaluation's results framework, can be invaluable. But just like quantitative graphics, their value depends on how well they are designed and explained.

As noted earlier, there are likely to be sensitivities surrounding the reporting of program results, particularly if the report contains bad news about program performance, program staff, key program functions, or something else. Given this, it is important, at least in the formal report itself, to present results tactfully but also as objectively as possible. If more pointed criticisms of program staff or functions are warranted, off-the-record discussions in staff meetings or other settings may be the most effective way of delivering such feedback.

For funders and policy audiences, it is also worth considering a highly condensed version (one or two crisp, bulleted paragraphs, with a visual) of your results that will fit on a webpage. This type of piece is sometimes referred to as "microcontent." Creating a piece like this is consistent with the layered communication approach we have been discussing in this chapter. The web piece can function as the top layer (that is, the least detailed), beneath which your summary is the next layer, and your report the final one.

14.3.6 Conclusions and recommendations

A **conclusions and recommendations** section is especially critical for audiences that you are calling on to do something and should also point the way to future work, if needed. It is important to make sure that the conclusions being presented are all logically derived from the findings that were in the previous section. Conclusions are typically at a higher level than findings, but they should not appear out of "thin air" in this section. A reader should be able to see the linkages or, better yet, the linkages between findings and conclusions should be explicitly called out.

The same is true for the recommendations. While not every conclusion may have an associated recommendation with it, every recommendation should have an associated finding(s) linked to it.

BOX 14.1 REPORTING RESULTS IN JOURNAL ARTICLES

Increasingly, it is also true that funders and policy makers expect program evaluations to publish their results. There may be many reasons for this expectation. For example, funders may want your evaluation's methods and results to be **peer-reviewed** as a check on quality assurance. Or they may be eager to avoid the "file drawer effect" discussed earlier. In such cases, even if the main audiences are not academic or research professionals, you will likely want to prepare a journal article based on your results and submit it to a scholarly journal in the relevant field.

In planning a journal article, you should be able to use the basic outline for the report provided above. But there are three likely differences between a report and an article. First, for the latter, you will need to detail your methods and data more carefully than you might have in a report intended only for policy audiences. Second, you are much more likely to face page limits or other formatting constraints in journal articles, and, thus, will have to report your results much more parsimoniously. Finally, most journal articles replace the "Conclusions and Recommendations" section of the report with a "Discussion" section that explores possible reasons for the results and considers next steps—whether these are recommendations, the need for further research, or other considerations.

Pointing the way to future work is also a key part of this section. If the formal report is documenting a formative evaluation, there are going to be "next steps" to point to that will be done before getting to the summative evaluation. But even if the report is documenting a summative evaluation, it can also point to future work that extends the currently reported work or positions the current work within the context of the field. Research builds up incrementally, and the work reported in the evaluation will be adding to the evidence base.

Make sure conclusions are logically derived from the findings and that the recommendations are logically derived from the conclusions.

14.4 DISSEMINATION: GETTING THE WORD OUT

Once you have a well-crafted report and any other products tailored to your key audiences, there is one more step: getting your findings in the hands of those audiences, as shown above in figure 14.1. Without this final step, the evaluation report alone will not have the desired effect, and important stakeholder groups that form part of your audience may need to be engaged in different ways. For instance, in an evaluation of a rural financial education program, findings may need to be shared with various different communities and their leaders, program officials at headquarters, local government officials and international donors—all of whom may focus on different aspects of the findings, and will need different modes of engagement.

We refer to these ways of engaging audiences as "channels." In today's media-rich technology world, it can be dizzying to try to catalog and track all possible dissemination modes. For the sake of a simple discussion, we discuss a simple typology for thinking about these.

Dissemination involves in-person, print distribution, and electronic modes.

- **In-person.** These can include informal meetings and discussion sessions, as well as more formal slide or multimedia presentations. These kinds of meetings may be an important way to engage internal audiences, funders, and/or policy audiences who might not otherwise turn to your report or your summary without some kind of in-person event to call it to their attention. In practical terms, it may be the only way to engage participants or communities in the field. While the evaluation team itself may not personally conduct meetings, this can be done by helping to provide materials and guidance to facilitate such discussions by program staff or local figures. Another important venue for in-person presentation is formal conferences where the results of evaluations like these can be presented, whether for academics or practitioners.

- **Print distribution.** Old-fashioned mailings of printed reports or independently printed summary documents may be the least effective way to reach key audiences, because nearly all these channels have been subsumed by their electronic cousins (see below). However, do not underestimate the power of having print copies to distribute in some contexts, including the in-person events mentioned above or at conferences where some of your audience members will be in attendance. Summaries or highlights of evaluation results may also be disseminated via existing print distribution channels such as local newspapers or organizational newsletters.

- **Electronic dissemination.** In some cases, electronic dissemination is understood rather vaguely to mean nothing more than uploading a PDF version of a print document to your organization's website. While this can be an important step in enhancing access to your evaluation results, electronic dissemination can involve a good deal more than this. If you want to bring audiences to your organization's website to view your evaluation results, you might consider creating an electronic newsletter that will do so. As noted above, in this context you may also wish to consider creating microcontent on your website as a high-level summary of your evaluation results and their implications for action.

- **Other media.** In addition to print and electronic dissemination, the evaluation team may wish to use other media platforms, e.g., interviews or features on local radio and television, in order to reach a wider audience. Where available, new social media outlets feeds are also increasingly important avenues for dissemination. For instance, several blogs of interest to the development community and evaluation in general are maintained by the World Bank and other organizations.

To bring all of these considerations together, you can create a dissemination plan that looks at the key audiences, specific products created to communicate with them, and channels through which you intend to reach these audiences. Table 14.2 is a sample dissemination plan for a program to educate the general public about financial capability. The secondary audiences include the other audiences identified in this chapter, but also the general public. The general public is an uncommon audience for evaluation research, but obviously an important audience for certain kinds of financial literacy education campaigns.

TABLE 14.2 SAMPLE DISSEMINATION PLAN

TYPE OF AUDIENCE	PRODUCTS AND ACTIVITIES	DISSEMINATION CHANNELS
Program participants	▪ Informal briefing	▪ Meetings or discussion groups ▪ Regular program communication channels
Program Staff/ Organization Upper Management	▪ Report ▪ Informal briefing	▪ Website (intranet) ▪ Meetings
Client	▪ Report ▪ Formal briefing ▪ Multimedia presentation	▪ Website ▪ Meetings
Practitioners (operators of or participants in similar programs)	▪ Report ▪ Journal article ▪ Electronic newsletters ▪ Briefings and podcasts ▪ Press release ▪ Webinars and webcasts	▪ Website (external) ▪ Annual conferences ▪ Email distribution list ▪ Smaller professional conferences, seminars, and workshops
Federal Policy makers	▪ Report ▪ Short policy briefs ▪ Electronic newsletters ▪ Briefings and podcasts ▪ Webinars and webcasts	▪ Website (external) ▪ Email distribution list ▪ Individual meetings, briefings, and testimonies before Congress and agency officials
General Public	▪ Report ▪ Short policy briefs ▪ Press release ▪ Electronic newsletters	▪ Website (external) ▪ Traditional media campaign ▪ Social media campaign ▪ Innovative channels developed in consultation with clients
Researchers	▪ Report ▪ Working paper and journal article	▪ Website (external) ▪ Presentations at academic conferences

14.5 PUTTING IT ALL TOGETHER

Communicating and disseminating your evaluation results begins, as shown earlier in figure 14.1, with identifying key audiences and their interests; it then involves documenting the program effectively and communicating what you want them to learn in appropriate ways that will engage their attention. Table 14.3 sums up the key points from this chapter, incorporating some of the information in table 14.2.

TABLE 14.3 KEY ISSUES IN COMMUNICATING AND DISSEMINATING EVALUATION RESULTS

AUDIENCE	WHAT THEY WANT TO LEARN	WHAT REPORT SHOULD EMPHASIZE	ADDITIONAL FORMS OF COMMUNICATION	DISSEMINATION CHANNELS
Program participants	How well is the program working? What works and what doesn't?	Program description, impacts	Discussion groups	Local community meetings or program communications
Program staff	How well is the program working? What works and what doesn't? What can we do better?	Program description, process results that highlight problems and how they can be addressed	Informal annotations to report for staff only or additional, informal internal communication	Staff meetings, internal discussions
High-level managers in own organization	How well is the program working? Should the program be continued, expanded, modified, or ended?	Bottom-line assessment of how well program is working; how results translate into action items	Slide presentations or talks; face-to-face meetings	Internal meetings, intranet posting
Funders	Is the program successful? Should the program be continued, expanded, modified, or ended?	Focus on summary; bottom-line assessment of how well program is working; how results translate into action items	Slide presentations or talks	Email newsletters; web posting; microcontent
Policy audiences	Is this type of program effective in addressing key problems?	Focus on summary; broad policy context; bottom-line assessment of how well program is working; how results translate into action items	Slide presentations or talks	Email newsletters, web posting, microcontent
Managers or staff in similar programs	What lessons can we draw to implement or improve our own programs?	Evaluation methods, program process, and implementation issues	Working paper; informal communication via professional networks	Your organization's website or professional website for professionals in the field
Researchers	What did this evaluation add to the evidence base for programs (or interventions) of this type?	Methods, validity of findings, contribution to evidence base in the field.	Journal article or working paper on website; conference presentations	Journal's website, links from your organization's website; paper or poster delivered at conferences

KEY POINTS

Documenting and communicating the results of evaluations is a critical last step and it is one that starts with carefully thinking about the logical audiences for the evaluation results and designing the document to best fit those needs.

Figuring out the best ways to disseminate the report—through in-person communications at meetings and conferences, traditional print means such as reports, and electronic venues such as websites and electronic newsletters—is also key to ensuring that the evaluation results are seen and used by the key audiences identified.

FURTHER READING

General

Bamberger, M., K. Mackay, and E. Ooi. 2005. *Influential Evaluations: Detailed Case Studies*, Washington, DC: World Bank.

Cronbach, M. 1980. *Toward Reform of Program Evaluation*, Jossey-Bass.

Kinneavy, J. 1971. *A Theory of Discourse*, New York and London: Norton Publishing Co.

Morris, L. L., C. T. Fitz-Gibbon, and M. R. Freeman. 1987. *How to Communicate Evaluation Findings*, Sage Publications.

Scargle, J. D. 2000. "Publication Bias: The 'File-Drawer' Problem in Scientific Inference." *Journal of Scientific Exploration* 14 (1): 91–106.

Technical

Baltussen, R., Ament, A., and Leidl, R. 1996. "Making Cost Assessments Based on RCTs More Eseful to Decision-Makers," *Health Policy* 37 (3): 163–83.

Appendixes

APPENDIX A

Trust Fund financial capability evaluation projects

A.1 THE IMPACT OF FINANCIAL LITERACY TRAINING FOR MIGRANTS: EVIDENCE FROM AUSTRALIA AND NEW ZEALAND

Remittance costs are high in the Pacific, which reduces the development potential of these flows. The efficacy of policies geared toward reducing the cost of remitting and spurring competition by increasing disclosure of costs relies heavily on the abilities of migrants to understand how to use the different methods available for remitting and the costs incurred for each method. Remittance-specific financial literacy training is being used in Australia and New Zealand to determine whether this will serve to induce migrants to choose lower-cost methods.

While systematic evidence on the financial literacy of migrants is scarce, the data available suggest migrants often lack knowledge of the components of a remittance cost, available methods, or how to compare such methods. The main goal of the present research is, therefore, to estimate the causal impact of financial literacy training for migrants on their remitting behavior. In particular, the research will assess whether (a) financial literacy training leads migrants to adopt new, cheaper, products such as debit cards for sending remittances; (b) whether financial literacy training changes the amount and frequency of remitting, and therefore the amount received by migrants in the home country; and (c) how these effects differ by country of origin.

Additionally, many migrants do not use credit cards and rely on more expensive forms of credit such as pay-day loans and hire purchase agreements. A second component of the financial literacy training will be to provide information on the costs of these alternative forms of credit, and information on how to apply to get a credit card, with the goal of seeing whether this reduces the use of more expensive forms of credit.

Short-term follow-up surveys were then conducted at one, two, and three months after training to track short-term impacts of the training on knowledge and behavior; with a follow-up after nine months scheduled. Experimental impacts of the program will then be obtained through comparison of the treatment and control groups.

A.2 FINANCIAL EDUCATION AND BEHAVIOR FORMATION: LARGE-SCALE EXPERIMENTAL EVIDENCE FROM BRAZIL

This study is the first large-scale, rigorous, impact evaluation to measure whether a high school-based financial education program can successfully change financial knowledge, attitudes, and behavior among students and their parents. The results of this study will inform the roll-out of the program to other public schools in Brazil. They will also be useful for other governments in developing countries that are considering the possibility of introducing financial education programs in schools.

Improving the financial capability of the population so that citizens are able to make effective decisions around personal finances is especially pertinent in Brazil, where the rapid evolution of the financial markets has resulted in many inexperienced and vulnerable consumers accessing different financial products and services, often with very negative consequences.

A 2008 survey found that 82 percent of Brazilian consumers were unaware of the interest rate when borrowing money, that overdue installments were mostly caused by poor financial management, and that the saving rate of Brazilians is low, even among affluent families. The survey also showed that 87 percent of families do not save for the future, and that 40 percent do not make any sort of investments with excess income.

This research project is divided into two parts: (1) measuring the impact of financial literacy programs for high school students on their financial knowledge acquisition and changes in financial decision making and behavior of their households; and (2) measuring the impact of a financial literacy program for parents of high school children. The behavior changes to be measured include changes in household financial attitudes and decisions, and subsequently changes in household's consumption, income, health and education expenditures. Information on financial knowledge, attitudes, and behaviors will be collected through surveys at the student and household levels.

A.3 THE IMPACT OF FINANCIAL EDUCATION AND LEARNING-BY-DOING ON HOUSEHOLD INVESTMENT BEHAVIOR: EVIDENCE FROM BRAZIL

The study offers a unique opportunity to study the individual and interaction effects of financial literacy campaigns together with multiple other interventions to better understand household investment behavior. While being applied in the Brazilian stock market, the lessons learned will help illuminate household financial decision making in general and offer suggestions on how to effectively use financial literacy campaigns to support consumer protection and promote a new generation of people accessing financial instruments in emerging markets

Understanding why people make the investment choices they do and what influence financial literacy can have in this decision process can help identify interventions that can both improve consumer protection and support the development of capital markets by improving efficiency through better investment decisions at the individual level.

Drawing from the current literature on behavioral finance, this study will work with the Brazilian stock market (BM&FBOVESPA) to test a range of interventions to better understand why people make the decisions they do and identify how to reduce common investor biases observed in the stock market through targeted financial education.

BM&FBOVESPA has developed an online stock market simulator that will be used as an environment to test barriers to entry for stock market participation while the "home broker" software provided by stock brokers in Brazil (which closely resemble the online simulator, offering investors the opportunity to invest online) will be used to selectively provide interventions to stock market participants to reduce investor biases.

A.4 INCREASING THE IMPACT OF CONDITIONAL CASH TRANSFER PROGRAMS THROUGH FINANCIAL LITERACY IN THE DOMINICAN REPUBLIC

This research project will evaluate the impact of financial literacy training offered to beneficiaries of the Solidaridad conditional cash transfer (CCT) program in the Dominican Republic, in conjunction with the Social Cabinet of the government of the Dominican Republic (GCPS).

Evaluations of CCT programs such as Oportunidades in Mexico and Familias en Accion in Colombia have shown that such programs are successful in increasing usage of health care and education services. However, even with improved health and education outcomes, it may remain difficult for CCT beneficiaries to manage their household finances, find stable employment, or start profitable businesses. All of these problems affect beneficiaries' ability to graduate from the CCT program and achieve a sufficient level of economic stability on their own. Previous research conducted by the DR government has shown that Solidaridad beneficiaries tend to have low levels of financial literacy and little access to financial services.

In this context, Solidaridad is piloting a series of projects to increase financial literacy and access to credit and savings to improve the income generating opportunities of current beneficiaries. This financial literacy program targets the heads of households in Solidaridad with a moderate poverty level, of which 40 percent are employed at least part time and over 85 percent of which have at least primary school education (SIUBEN 2005).

IPA will use a randomized controlled trial (RCT) design to evaluate the impact of the training programs. In an RCT, one group of beneficiaries is randomly selected to participate in the training (the "treatment" group), while another is randomly selected not to receive the training at the time of the evaluation (the "control" group). Randomly assigning beneficiaries to treatment or control allows us to ensure that any differences observed between the two groups after the training program is over can be directly attributed to the training. The package includes training in household and business financial management, job skills, and access to financial products. The training program is expected to have short-term effects on beneficiaries' behavior, and long-term effects on their overall welfare.

A.5 DOES FINANCIAL EDUCATION AFFECT SAVINGS BEHAVIOR? EVIDENCE FROM A RANDOMIZED EXPERIMENT AMONG LOW-INCOME CLIENTS OF BRANCHLESS BANKING IN INDIA

In partnership with the World Bank evaluation team, FINO, a "doorstep banking" firm, has developed and implemented a pilot financial education program to support the increased use of FINO's smart card as a mechanism to encourage and facilitate saving. The financial education program focuses on teaching the knowledge and skills required to adopt good money management practices for budgeting, spending, and saving. Participants in the financial education program are expect to be equipped with the information and tools necessary to make better financial choices, work toward their financial goals, and enhance their economic well-being.

This study is aimed at measuring the impact of the FINO financial education program on the participants' knowledge, attitudes and behavior related to personal financial management and overall financial well-being. One of the innovations of the FINO financial education program is that it will examine how people interact with the technology that facilitates their access to bank services, in this case, the FINO smart card and the Point of Transaction (POT) of the Business Correspondents (BCs). Research shows that one of the barriers to the uptake and use of branchless banking in many developing countries is the low levels of familiarity and trust with the technology behind electronic cards and mobile phone banking among the poor. The FINO financial education program will therefore seek to add to individual's financial literacy as well as address the gaps in the operational knowledge and skills in conducting smart card transactions, as well as strengthen trust.

This evaluation utilizes a randomized treatment and control identification strategy. For the purposes of this study, the treatment group is defined as those individuals living in villages where the training program will be implemented in the coming months, while the control group is villages which will receive training during a later phase of the program, after the last follow-up survey has been conducted. If the treatment and control are balanced at baseline, then differences at follow-up for key

indicators can be attributed to the intervention(s), rather than to some preexisting differences between the two groups. A lottery system was used to select which trained BCs provided financial literacy training.

The present research ultimately looks to develop a more precise understanding of the financial literacy levels of the low-income populations of India and the extent to which that is a barrier and/or facilitator of their personal financial management in conjunction with doorstep banking. The results are expected to inform researchers and policy makers regarding whether providing financial education in conjunction with financial access, all developed in an environment that is familiar and comfortable for the participants, is an effective approach to improving savings and inducing sound financial management among the unbanked poor.

A.6 THE ROLE OF FINANCIAL ACCESS, KNOWLEDGE, AND SERVICE DELIVERY IN SAVINGS BEHAVIOR: EVIDENCE FROM A RANDOMIZED EXPERIMENT IN INDIA

A new financial inclusion and literacy impact evaluation will be conducted jointly with Eko India Financial Services, a microfinance institution in India, in order to expand the research base on the impact that knowledge of financial services can have on financial literacy and financial inclusion. It will also deepen the learning on how various financial services delivery modes impact upon the overall welfare of low-income households and on their savings behaviors.

Eko is one of the few institutions to offer "doorstep banking" and financial services in this area of India. Most of the villages in which Eko operates have little or no access to banks. The problem of access to banking is further exacerbated by the fact that the people with whom Eko works come from very low income families and thus face numerous difficulties in opening savings accounts. To solve these access issues, Eko utilizes a savings program in the villages. Thus, people can conduct immediate transactions within their village instead of traveling long distances to access a bank.

In terms of the project's evaluation methodology, the treatment group will be given financial literacy education, particularly about the importance and benefits of savings. Following the training, researchers will present simple and cheap methods for follow-ups to increase the retention of information from the training. The program is a multiround, financial education intervention where an initial round of training will be provided in a classroom session using standard tools, which will be reinforced by intensive follow-ups. The treatment group will receive additional rounds of training via their mobile phones through SMS and voice messages for the first few months, and will be physically visited by a trainer after six months.

From this project promoting increased access to doorstep banking, researchers expect to see an increase in the number of financial transactions customers under-

take. Meanwhile, the training component will help shed light on the mechanisms through which intra-household financial decisions are made. Finally, researchers will explore the potential that intensive financial education training has to influence financial behaviors and well-being by comparing treatment with control groups.

A.7 THE IMPACT OF CARTOONS AND COMICS ON THE EFFECTIVENESS OF FINANCIAL EDUCATION: EVIDENCE FROM A RANDOMIZED EXPERIMENT IN KENYA

Kenyan youth face an uncertain and volatile financial landscape, with high un- and underemployment, questionable long-term job prospects, and little or no protection against the vagaries of ill health and injury. On the other hand, however, young Kenyans have more opportunities to invest in and plan for their futures than perhaps any earlier generation: market liberalization and a stable macro-economy have facilitated steady growth in recent years, educational and training options abound, and access to financial services has expanded considerably.

The study is being implemented by two Kenyan organizations that have sought to help young Kenyans prepare themselves for economic opportunities through innovative means. It looks to identify the effectiveness of alternative modes of delivering financial education to high school students. It compares the standard delivery of a structured course of materials with a series of weekly comic book episodes that personalize, contextualize, and make pertinent to the target population the lessons of the course.

The authors combine a randomized control trial with quasi-experimental quantitative techniques to assess the impact of each of the interventions, compared with no program, as well as vis-à-vis each other. The interventions are randomized across Junior Achievement Clubs.

In order to measure the impact of the interventions, the authors will collect data at baseline, before the programs, and at endline after their completion. As well as administering a survey to about 5,000 students, they will engineer a situation in which a nontrivial financial decision must be made. First, both at baseline and endline all students are asked how they would allocate about $25 of unexpected earnings between cash, a bank account, and a mutual fund on the Nairobi stock exchange. Subsequently, authors deliver prizes of this amount to each of roughly 2,000 students, again at baseline and endline. The hope is to discern changes in both the stated and the actual allocations, between baseline and endline in the treatment groups. This will provide a unique outcome measure of students' actual financial behavior, to complement the more qualitative data we collect in interviews.

Finally, a series of behavioral games will be conducted to measure differences in attitudes and preferences, for example related to discounting, patience, and long-term perspectives across students exposed to the different treatments.

Results from the study are expected to offer unique insights into how best to reach and teach young people. To the extent that sound investment, careful debt management, and prudent saving and business decisions are part of a successful transition over the longer term to middle-income country status, understanding the power of mass communication channels in inducing behavior change is of interest to policy makers in East Africa and beyond.

A.8 SOCIAL NETWORKS, FINANCIAL LITERACY, AND INDEX INSURANCE: EVIDENCE FROM A RANDOMIZED EXPERIMENT IN KENYA

Regarded as a promising alternative to traditional crop insurance, market participants, NGOs, donors, and governments are all testing the applicability of index based weather risk management instruments in different contexts. A significant advantage of this type of product is that payouts can be calculated and disbursed quickly and automatically without the need for households to formally file a claim because they are based on measured rainfall. This, in turn, reduces transaction costs, which would otherwise tend to drive up the price of the insurance. Fast payouts are also likely to be valued by policyholders in an environment where households are poor and often liquidity-constrained. A second advantage is that the product is free of adverse selection and moral hazard problems that often plague insurance markets. This is because payouts are based only on publicly observed data, rather than private information about the beneficiary.

The goal of this study is to examine the role of financial literacy on farmers' decisions to purchase index-based weather insurance and to examine social network spillovers of financial literacy provision in Kenya. The researchers propose to benchmark these effects with the effect of providing discount vouchers off the price of insurance.

The study uses comics as the central financial literacy delivery mechanism. The comic used in this study describes a family which had faced a drought in the previous season. The drought had adverse effects on their savings and well-being. The comic carefully details the index-insurance product and shows how it can help the family protect themselves from the risk of drought. The comic presents the insurance product in an accessible and relevant manner and has sustainability as an educational tool as it can further be shared with other farmers.

The study employs a randomized controlled trial (RCT) to determine which type of financial literacy is most effective. Two interventions were tested: a comic on index insurance and discount vouchers for the purchase of insurance. At the cluster level treatment intensity was randomized and orthogonal across treatments. This

design allowed for an accurate assessment of spillover effects of financial literacy compared to spillover effects of higher participation from the discount vouchers. After completing the baseline survey, enumerators administered the interventions. Each household was randomly assigned to receive a comic or not depending on what comic intensity the household's cluster was assigned.

A.9 FINANCIAL DEVELOPMENT AND THE PSYCHOLOGY OF SAVINGS: EVIDENCE FROM A RANDOMIZED STUDY IN MALAWI

This project investigates innovative ways to address low levels of formal savings in developing countries. The study explores to what extent psychological mechanisms can be leveraged to increase formal savings. In particular, the study determines whether direct deposit of wages—as opposed to receiving cash—can help individuals to match desired savings and expenditure patterns with actual behavior. In addition, the study explores how a combination of formal financial products and training can help to activate mental accounting to facilitate savings.

The study has two parts: In the first part, it evaluates the introduction of a new direct deposit system at large agricultural firm in Malawi that pays workers' wages into individual accounts, instead of paying in cash. Receiving wage or farm revenues in the form of cash may exacerbate self-control problems, since the temptation to spend may be higher when cash is on hand. Drawing on surveys and administrative data from bank records and the tea estate the study seeks to measure changes in the consumption patterns—timing and allocation across expenditure categories– as well as formal and informal savings and borrowing behavior.

The evaluation utilizes a randomized controlled trial approach. As described above, the project consists of two parts: the first evaluating the impact of switching from cash to direct deposit, the second exploring the usefulness of labeled accounts to activate mental accounting In the first part, working with a population of wage earners at a large agricultural firm in Malawi, a new direct deposit system of wage payments will be introduced using a random phasein. A subset of workers will be switched to the new system earlier while the remainder serves as a control group to be able to compare financial behavior and well-being between the two groups. A complementary, cross-randomized roll-out of training sessions with individual households will evaluate whether impacts can be amplified through financial literacy training, and whether impacts of the introduction of direct deposit for wages can be emulated with such training.

In the second part, the project looks to study mental accounting and the role of labeled bank accounts to activate such mental accounting. Mental accounting describes the phenomenon that money is not always fungible across expenditure categories. Savings that are mentally assigned to a specific savings goal or set of

expenditures, like inputs for farming or children's school fees, may become less available for expenditures for another set of expenditures, like unplanned purchases. Offerings savings accounts that stress such mental separation by being labeled for specific expenditure categories such could therefore help individuals achieve their savings goals. The study seeks to measure impacts on financial behavior and household well-being through combination of survey and administrative records. Additional behavioral experiments will help to disentangle competing theories about effect channels and the nature of mental accounting.

A.10 EVALUATING THE EFFECTIVENESS OF LOAN DISCLOSURE REFORMS ON CONSUMER UNDERSTANDING AND FINANCIAL DECISION MAKING: EVIDENCE FROM MEXICO

An important question in financial capability policy is whether consumers learn financial lessons more easily through school-based traditional methods (such as in the classroom, in training seminars and in workshops, etc.), or by accessing clear and easy to understand information in the marketplace. In Mexico, this question is of increasing importance for a number of reasons: increased use of financial products in recent years (particularly among low-income consumers); evidence of repayment problems and hidden charges in some credit and savings products; and many types of regulated and nonregulated financial institutions, making oversight and market monitoring a constant challenge.

This study explores the intersection between consumer protection and financial capability by comparing how product-specific information and broader financial capability messages impact short and medium-term financial decision making of the low-income consumers for basic credit and savings products. It will also evaluate the impact of several different channels for communicating important financial messages to consumers, including: point-of-sale disclosure; information to facilitate comparison shopping; financial counseling by phone; and periodic financial capability SMS messages.

By exploring this intersection the research hopes to offer new insights into whether the type of market conduct reforms and policy interventions that have been used to date solely for consumer protection purposes can actually be used to improve financial capability and influence sound financial decision making. With the high cost of traditional curriculum-based financial education programs, finding regulatory approaches to improve financial capability has the potential to offer promising policy options for resource-constrained governments interested in improving financial capability levels.

A.11 MEASURING THE IMPACT OF FINANCIAL LITERACY ON SAVINGS AND RESPONSIBLE CREDIT CARD USE: EVIDENCE FROM A RANDOMIZED EXPERIMENT IN MEXICO

The study tested the impact of financial literacy training on savings and borrowing behavior and credit card usage patterns of credit card customers in Mexico City. It involved approximately 40,000 bank consumers. The training course lasted for about four hours and consisted of four modules on savings, retirement, credit cards, and responsible use of credit. Results show a 9 percentage point increase in financial knowledge, and a 9 percentage point increase in saving outcomes, but no impact on credit card behavior, retirement savings, or borrowing. Moreover, administrative data suggest that the savings impact is relatively short-lived. The results point to the limits of using general-purpose workshops to improve financial literacy and decision-making patterns for the general population.

A.12 THE IMPACT OF FINANCIAL LITERACY THROUGH FEATURE FILMS: EVIDENCE FROM A RANDOMIZED EXPERIMENT IN NIGERIA

Low-income households often lack access to formal education. Using popular culture and entertainment as an alternative delivery mechanism to promote financial literacy is a potentially powerful means of facilitating access to financial information. This research project evaluates the extent to which a feature film produced through the Nigerian Film Industry (Nollywood) can increase financial capabilities of households by promoting responsible borrowing and saving strategies.

This research looks to provide inputs to the current debate in the financial capability field around of the value and viability of non–classroom style, alternative delivery mechanisms. Leveraging the support of Nollywood, the initiative uses a full-feature film, *The Story of Gold*, to enhance financial literacy in viewers. It then assesses the impact of the film on individuals' financial capabilities (namely, literacy, skills, attitudes, and behavior).

Essentially, it looks to test whether a popular media initiative such as this is able to: (1) increase people's awareness and understanding of best practices for day-to-day financial management, saving and borrowing money; and (2) lead to more responsible financial decision making. The movie aims to teach low-income individuals with limited formal education some of the core concepts around financial planning. The film revolves around the core values of "smart savings and responsible borrowing" and disseminates this message through large screenings in public locations across Lagos State. Focusing on this simple behavioral goal and highlighting the repercussions of poor financial decisions, *The Story of Gold* looks to leverage the popularity of Nollywood to build awareness and impact the financial decision-making processes of low-income Nigerians.

Markets will be selected in the outskirts of Lagos and a screening area will be chosen that will be accessible for all participants in order to maximize take up rates. An initial listing will be conducted that will entail a full census of stall owners within a 1 mile radius of the screening event. The listing will act as a mini baseline and will include a selection of key indicators of interest together with identification indicators (including GPS coordinates). Once the listing has been completed, participants will be randomly selected into four different groups: (1) *The Story of Gold* screening only, (2) *The Story of Gold* screening with MFIs opening savings accounts for viewers on the spot, (3) a placebo movie screening (control) and (4) the placebo screening with MFIs. Survey participants will receive color-coded wrist bands and personal invitations the day prior to the movie screening to minimize the chance of contamination and to improve participation rates. By randomly selecting who will receive which intervention, we will be able to measure the causal impact of both the movie screening and the movie screening reinforced with MFI representation to disentangle the long-term educational effects of the movie from the immediate emotive context on influencing savings and borrowing behavior.

A.13 LEARNING BY DOING? USING SAVINGS LOTTERIES AND SOCIAL MARKETING TO PROMOTE FINANCIAL INCLUSION: EVIDENCE FROM AN EXPERIMENT IN NIGERIA

This project explores how new media and "learning-by-doing" can encourage financially unsophisticated consumers to open and maintain savings accounts. Researchers will evaluate a promotion by a large Nigerian bank to examine (1) how consumers respond to different types of new media campaigns, e.g., internet and celebrity endorsements, and (2) how the experience of maintaining a savings account over a three month period can improve financial literacy and change long-term precautionary savings habits.

In 2011, InterContinental Bank in Nigeria launched a nationwide savings promotion called "I-Save I-Win" (ISIS). It featured a large number of heavily publicized lottery prizes for those who opened or maintained savings account and maintained savings balances above various threshold amounts—N50,000 ($320) for regional lotteries, N100,000 ($640) for the national prize—for 90 days. Its message: although the special promotion included lottery prizes, every saver is a "winner." ISIW was promoted with a media push including celebrity endorsements and media releases through Facebook and YouTube, which were staggered over a period of several months.

This project will evaluate ISIW's impact on customers' savings habits and financial literacy. First, researchers will explore how the experience of maintaining savings accounts during the 90-day promotional period affects savers' long-term savings habits. Second, they will examine how different media promotions affect customer participation in the program and savings behavior: which media pushes led to more

sign-ups, and to what extent did each lottery winner have a "demonstration effect," i.e., inspiring others to increase savings at that particular branch.

To study the extensive margin of how ISIW affected new account sign-ups and savings, we conduct a similar differences-in-differences analysis comparing sign-ups at InterContinental Bank before and after each phase of the savings promotion, and again control for time trends using data from a second bank that did not have a savings promotion.

In addition, we examine the effect of winning the lottery on lottery winners' savings behavior and the savings behaviors of individuals associated with the same banking branch as the lottery winner. The lottery randomization offers a clean source of identification through which we can establish a causal effect of lottery winnings on savings behavior.

We will also conduct surveys to learn about changes in financial behaviors that are not directly observable from bank account data, such as financial literacy and the source of funds brought into InterContinental Bank (e.g., consumption, other banks, stored cash).

A.14 HARNESSING EMOTIONAL CONNECTIONS TO IMPROVE FINANCIAL DECISIONS: EVALUATING THE IMPACT OF FINANCIAL EDUCATION IN MAINSTREAM MEDIA IN SOUTH AFRICA

This project aims to pilot entertainment education as an information delivery tool to influence sound financial management in South Africa. It will do so by assessing the usefulness of this instrument to increase financial capabilities of the population with a robust impact evaluation framework.

The goal of the intervention is to impact upon the knowledge, attitudes and behavior regarding sound financial decision making with a particular focus on managing debt. The potential gains of such a program in terms of increased awareness of debt-related problems and behavior change, such as avoiding debt and seeking help once over-indebted, are of particular relevance in South Africa given the high and increasing levels of household debt.

The project entails the content development, production, and impact evaluation of financial capability storylines, which will be included in a popular South African soap opera called *Scandal!*. The target individuals are low-income South Africans with and without existing consumer debt. The financial education storyline will stretc.h over a period of three consecutive months, a time that is deemed necessary for the viewers to emotionally connect with the soap opera characters, for the events to unfold, and for the financial literacy messages to sink-in.

The estimation of the impact of the soap opera on knowledge, attitudes, and behavior will be straightforward as randomization allows for a direct comparison between the treatment and the control group. The random selection and comparison to a control group allows for attribution of the effects found. The financial incentive insures that a comparison sample is available for interviews and the incentive for both treatment and control groups eliminates any confounding effects of the incentive itself. Cumulative and/or nonlinear effects of the financial literacy messages can be detected through several rounds of interviews. And lastly, spillover effects can be measured by comparing those who did not watch but heard about the content with those who neither watched nor heard about it.

A.15 THE IMPACT OF FINANCIAL EDUCATION ON FINANCIAL KNOWLEDGE, BEHAVIOR, AND OUTCOMES: EVIDENCE FROM A RANDOMIZED EXPERIMENT IN SOUTH AFRICA

Governments, NGOs, and aid organizations are increasingly focusing on financial literacy education as a tool for improving welfare. Yet to date, there is little rigorous evidence that financial education is effective. This project evaluates Old Mutual's "On the Money," a one-day financial education program that provides training on saving, financial planning, budgeting, and debt management. The training program is very similar in content and delivery to a financial literacy evaluation being conducted by the same authors in India.

To rigorously measure the impact of the intervention, this project use a randomized control trial involving approximately 1,300 individuals: 610 organized in 43 Burial Societies and 690 organized in 36 Women's Development groups in the Eastern Cape and KwaZulu Natal. A randomly selected half of these groups receive financial education. The other half forms the control group for the duration of the study.

To estimate the true causal impact of financial education, it is critical to establish the correct counterfactual. Studies that simply compare individuals who receive financial education to those who do not are susceptible to selection bias, meaning that people who choose to take financial literacy courses may differ from those who choose not to take such courses. The RCT evaluation methodology eliminates that bias. By randomizing assignment to treatment or control, researchers ensure the offer to attend training is not correlated with any potentially confounding factors like level of education, income, or motivation.

Examining the scope and pathways by which financial education affects financial behaviors requires that researchers obtain a rich data set on trained participants and the control group. To that end, we will obtain outcome measures from a variety of sources, including individual surveys and administrative data collected by Old Mutual and WDB, which will allow for an understanding of how financial literacy evolves, the ways individuals change their financial behavior in response to training.

A.16 UNDERSTANDING FINANCIAL CAPABILITY THROUGH FINANCIAL DIARIES AND IN-DEPTH INTERVIEWS IN UGANDA

The Understanding Financial Capability through Financial Diaries and In-Depth Interviews in Uganda project explores the use of an innovative survey methodology—Financial Diaries—in combination with in-depth interviews, to better understand and measure the financial capabilities of low-income individuals. Financial Diaries are multiperiod surveys of individuals that record all their economic transactions over a period of a number of months. This is a rich source of detailed information on the economic behavior of respondents, including their financial service use. The diaries are thought to be a potentially useful tool for analyzing behavioral change following participation in financial education programs and other initiatives. To examine that potential, the project addresses the following three questions:

- What can Diaries, in combination with in-depth interviews, tell us about the financial capabilities of low-income people that we might not know otherwise?
- How can change in indicators of financial capability be tracked through Diaries, in combination with in-depth interviews, over time?
- Under what circumstances is it appropriate to use financial Diaries to evaluate the impact of a financial education program?

The project involved 103 respondents, 47 in a treatment group in two communities where Habitat for Humanity Uganda (HFHU) is offering financial education and 56 in a comparison group outside of HFHU's service area. All respondents reside in the Luwero district and in economically and demographically similar communities to each other. Financial diaries data are collected at weekly visits with respondents before, during and after the intervention. The period after the intervention consists of two time periods: one immediately following the intervention and another 12 months after the start of the Diaries data collection. Researchers will thus observe changes in behavior (manifested in changes in the pattern of economic transactions) made in both the short- and the long-term in both the treatment and comparison groups. In-depth interviews are conducted with respondents before, immediately after, and 10 months after the intervention to identify any changes in knowledge among the treatment and comparison groups.

A.17 THE IMPACT AND NETWORK EFFECTS OF FINANCIAL MANAGEMENT AND VOCATIONAL TRAINING IN INFORMAL INDUSTRIAL CLUSTERS: EVIDENCE FROM UGANDA

Identifying the determinants of entrepreneurship is an important research and policy goal, especially in emerging market economies where lack of capital and supporting infrastructure often imposes stringent constraints on business growth.

Studying how business networks operate is elemental for smart policy making as any identified positive externalities could justify scaling down spending on business and financial training programs while still being able to reap many if not all the benefits. These potential positive spillovers would constitute efficient ways to scale the impact of trainings and provide a natural source of leverage for these programs. Furthermore, the resources saved in doing so could be spent on expanding training in other areas or on other development projects.

The impact evaluation uses a randomized control trial (RCT) to assess whether vocational and business training for small scale entrepreneurs can impact business knowledge and outcomes. This is a competitive market of small-scale producers (metal fabricators, shoe makers, caterers, and the like) operating in industrial clusters in the outskirts of Kampala, Uganda. The research maps out business networks and seeks to identify to what extent such enhanced knowledge is shared among network members.

After the trainings are implemented, we will conduct follow-up surveys not only on treatment and control clusters (in order to get first-order effects of our training program), but also on network members of both treatment and control groups. The comparison of information sharing and outcomes for these groups will then identify the value of business networks.

APPENDIX B

Technical appendix

Chapter 6 (Impact Evaluation) introduces a full menu of quantitative methods that are available for impact evaluation. The purpose of this technical appendix is to provide a brief methodological overview of the methodologies described in chapter 6. For details on how to implement a particular evaluation methodology, please consult the references listed at the end of this appendix.

B.1 CAUSAL INFERENCE

We are interested in understanding the causal effect of a financial capability program on financial knowledge and behaviors. In order to understand how a program affects outcomes, we need to be able to answer the counterfactual question: "What would have been the financial outcome of the individual who participated if they did not participate?" Since an individual can either be exposed to the program or not, we cannot obtain an estimate of the program effect for an individual. However, we can obtain the average effect of the program on a group of individuals (the treatment group) by comparing them to outcomes of similar individuals who were not exposed to the program (the comparison group). As discussed in chapter 6, because any difference in outcomes between the treatment and comparison groups will be attributed to the program, it is important to control for preexisting differences between the two groups. All of the methods that were described in detail in chapter 6 control for differences in the two groups in various ways.

Using mathematical notation and following Duflo et al. (2008), let Y_i^T denote the outcome of individual *i* if they receive the financial capability intervention, and let Y_i^C denote the outcome of that same individual if they do not receive the intervention. Also, let Y_i denote the actual outcome of individual *i*. We are interested in the effect of the financial capability intervention: $Y_i^T - Y_i^C$. Because this is unobserved for any individual since they can only be in the treatment or comparison group, we cannot calculate this difference. However, we can calculate the expected average effect of the financial capability program: $E[Y_i^T - Y_i^C]$, where $E[\cdot]$ stands for the expectation function.

We could collect observational data on individuals who receive the intervention and those who do not. The difference in outcomes for a large sample of individuals can then be written as:

$$D = E\,[Y_i^T|T] - E\,[Y_i^C|C]$$

where the first term is the outcome after receiving the intervention, conditional on receiving the intervention, and the second term is the outcome without the intervention, conditional on not receiving the intervention. Notice that we can add and subtract the term $E\,[Y_i^C|T]$, and then this expression can be rewritten as

$$D = E\,[Y_i^T|T] - E\,[Y_i^C|T] - E\,[Y_i^C|C] + E\,[Y_i^C|T] = E\,[Y_i^T - Y_i^C|T] + E\,[Y_i^C|T] - E\,[Y_i^C|C]$$

The first term is the treatment effect, the difference in outcomes that we are interested in estimating. The second and third terms together measure the selection bias—they capture the difference in potential untreated outcomes between the treatment and comparison groups. The aim of the impact evaluation methods discussed in this Toolkit is to eliminate the selection bias or use statistical methods to correct for it.

B.2 RANDOMIZED CONTROL TRIAL

One method to remove selection bias is to randomly assign individuals to the treatment and comparison groups. The average treatment effect can be estimated using a large sample of treatment and comparison groups. The difference in outcomes is given by:

$$\hat{D} = \hat{E}\,[Y_i|T] - \hat{E}\,[Y_i|C]$$

where $\hat{E}$ denotes the sample average (and in general $\hat{\cdot}$ refers to an estimate). Because the financial capability intervention is randomly assigned, in an ideal setting the treatment and comparison group are identical on average, and the selection bias term is equal to zero by definition.

The intervention effect under a randomized control trial can be estimated using linear regression methods. Define T to be a dummy variable (taking values of 0 or 1) for whether the individual is assigned to the treatment group. Then the regression equation of interest is:

$$Y_i = \alpha + \beta\, T_i + \varepsilon$$

where β is a mean zero, constant variance, independent and identically distributed error term. Estimating this regression with ordinary least squares, $\hat{\beta}_{OLS}$ gives the average treatment effect, an unbiased estimate of the financial capability program effect. This equation can be augmented to include a flexible function of a vector of explanatory variables $\boldsymbol{X}_i$.

B.3 ENCOURAGEMENT DESIGN AND INSTRUMENTAL VARIABLES

In cases where the random assignment of the treatment is not possible, it may be possible to randomly encourage some eligible participants to take part in the financial capability intervention. Let *Z* be a dummy variable for whether the individual receives the encouragement, while *T* remains the indicator for whether the individual participates in the financial capability program. In this case the Intent to Treat Effect (discussed in chapter 6, box 6.6 on treatment effects) is the difference in outcomes between the group that received the encouragement and the group that did not receive the encouragement:

$$D_{ITT} = E\,[Y_i|Z = 1] - E\,[Y_i|Z = 0]$$

In order to estimate the ITT program effect in a regression framework, you will estimate the following equation:

$$Y_i = \alpha + \beta_{ITT} Z_i + \varepsilon$$

where $\hat{\beta}_{ITT}$ is the coefficient of interest, and *Z* is an indicator for whether the respondent received the encouragement.

However, as described in chapter 6, in many settings policy makers are interested in the impact of the intervention itself. In the context of encouragement design, we are able to estimate the local average treatment effect (LATE), which is given by

$$\hat{\beta}_{LATE} = \frac{E[Y_1 \mid Z = 1] - E[Y_1 \mid Z = 0]}{E[T \mid Z = 1] - E[T_1 \mid Z = 0]}$$

The LATE estimate adjusts the ITT estimate for the difference in the fraction of individuals who participate in the financial capability program as a result of being encouraged to do so (first term in the denominator) and the fraction of individuals who participate even when they are not encouraged (second term in the denominator). The LATE estimator is the effect of the intervention on those whose treatment status was affected by the encouragement.

There are two underlying assumptions that identify the LATE estimator:

1. **Independence:** in this context, the independence assumption means two things
   - Receiving the encouragement identifies the causal effect of the program
   - Financial outcomes are not directly influenced by the encouragement
2. **Monotonicity:** that no person is discouraged to take up the intervention after receiving the encouragement

These assumptions should be examined on a case-by-case basis to ensure that the LATE estimate is in fact identified. For example, the independence assumption is violated if the encouragement is in the form of a cash incentive, and the outcome of interest is the amount of savings—in this case the cash incentive itself can affect the amount of savings, regardless of the financial capability intervention. Similarly, the monotonicity assumption is violated if the treatment group participants believe that the encouragement is a scam, and therefore decide that it is not safe to take part in the financial capability program.

In a regression context, the LATE estimate is identified by estimating the following system of equations:

$$Y_i = \alpha + \beta T_i + \varepsilon$$
$$T_i = \delta = \gamma Z_i + \nu$$

These equations can be estimated in two steps (using two stage least squares), where the predicted treatment status (from estimating the second equation) replaces the real treatment status in the first equation. In this case the standard errors of the first equation need to be adjusted for the two step process. Then β gives the LATE estimate of the intervention. Note that the process described here for obtaining LATE estimator is valid for any setting where there is an instrumental variable that influences the probability of treatment.

B.4 REGRESSION DISCONTINUITY DESIGN

In the case of a regression discontinuity design, the probability of assignment to the treatment group is a discontinuous function of one or more variables. This means that

$$E\,[Y_i^C|T, R < \bar{R} + \varepsilon, R > \bar{R} + \varepsilon] = E\,\mathrm{f}Y_i^C|C, R < \bar{R} + \varepsilon, R > \bar{R} + \varepsilon]$$

where R is the ranked assignment variable, and $\bar{R}$ is the threshold for the assignment variable and ε is a mean zero i.i.d. error term. For multiple assignment variables, it is possible to replace R with $f(\boldsymbol{R})$, where $f(\cdot)$ is any function of assignment variables $\boldsymbol{R}$. The underlying assumption for regression discontinuity is that within a small threshold of the assignment variable, the selection bias is zero. The effect of the intervention is estimated by comparing the outcomes of individuals on one side of the threshold with the outcomes of individuals on the other side of the threshold serving as the comparison group.

A regression discontinuity analysis should always start with a graphical representation of the outcome variable and the assignment variable. Typically the assignment variable is divided into bins, with two separate bins around the threshold value, and the average values of the assignment variable are plotted against the outcome vari-

able. There should be a clear jump in the outcome variable at the threshold value. This graphical representation can help inform the functional form for the regression and give a rough estimate of the program impact by comparing the outcomes before and after the threshold value. It can also show whether there are unexpected jumps in the outcome variable at points away from the threshold. If these unexpected jumps cannot be explained, it calls into question whether the discontinuity is believable.

The size of the bins (called the bandwidth) is an important choice, because it defines what is considered "close enough" to the threshold. There is a long literature on how to choose the bandwidth, and in the context of regression discontinuity it is discussed in Lee and Lemieux (2009). They suggest a method known as cross-validation, also known as the leave out procedure. Alternative methods should also be considered to ensure that the bandwidth choice is not driving the results.

To estimate the impact of the program effect under a regression discontinuity design setting, the equation is given by

$$Y_i = \alpha + \beta_{RD} T_i + \gamma R_i + \varepsilon$$

where T_i indicates whether the individual received the financial capability intervention and R_i indicates the value of the assignment variable. In general, while this equation can be estimated parametrically, because the misspecification of the linear function can bias the impact estimate β_{RD} and because the purpose of this method is to estimate a parameter at a cut point, a number of researchers argue for estimating the equation nonparametrically. Nonparametric methods include kernel regressions using a rectangular kernel function or estimating the model with local linear regressions. These methods are explained in detail in Lee and Lemieux (2009).

In practice, a simple way to implement the regression discontinuity design is to estimate two separate regressions on each side of the threshold. It is convenient to subtract the cut point from the assignment variable, and transform R to $R–\bar{R}$. In this case, the difference in the intercepts of the two equations will give the estimate of the program impact. It is generally good practice to allow the slope of the regression on the two sides of the threshold to differ. To the left of the threshold the regression equation is

$$Y = \alpha_l + \gamma_l (R - \bar{R}) + \varepsilon$$

and to the right of the threshold the regression is

$$Y = \alpha_r + \gamma_r (R - \bar{R}) + \varepsilon$$

Then $(\alpha_l - \alpha_r)$ is the estimate of the impact of the financial capability program. These models should be estimated for values close to the cutoff as determined by the bandwidth choice.

See Lee and Lemieux (2009) for extensions of this basic method to "fuzzy" regression discontinuity (where the probability of treatment at the threshold are not exactly 0 or 1), and how to calculate the standard errors of the estimates.

B.5 PROPENSITY SCORE MATCHING

Propensity Score models attempt to overcome the issue of selection bias using observable covariates by creating a single variable—a propensity score–that captures how differences in these covariates contribute to a eligible unit's probability of receiving the treatment. Next, the propensity score is used to match treatment and comparison individuals, and outcomes are compared between these groups to estimate the impact of the intervention.

The propensity score matching method is implemented in five steps:

1. Estimate the probability of participating in the treatment conditional on observed data (propensity score)
2. Match each participant to one or more nonparticipants using the propensity score
3. Examine and evaluate the balance in observed covariates across treatment and comparison groups
4. Estimate the difference in outcomes between the balanced groups using multivariate regression
5. Perform a sensitivity analysis

The propensity score is the probability of being in the treatment group conditional on observed measures:

$$p(x) = P(T = 1|X = x)$$

This equation is typically estimated by using a logit or probit model to predict the probability of participating in the treatment (the propensity score). According to Austin (2006), it is recommended that a rich set of covariates that are related to the outcome are used in the estimation, but variables that are only related to the treatment are not included because including these variables reduces the ability to form a match. In addition, it is recommended that an expansive use of transformations and interaction terms are included in the covariates (such as two- and three-way interactions and log transformations).

According to Rubin (2007), variables that should not be included are ones that "…are effectively known to have no possible connection to the outcomes, such as random numbers…, or the weather half-way around the world."

The matching of the treatment and comparison group participants can be accomplished in numerous ways, including nearest neighbor matching (matching each treatment group member with a comparison group member who has the closest propensity score), caliper matching (matching multiple comparison group members with in a range of propensity scores, mahalanobis metric matching (matching on the basis of Mahanalobis distance between subjects), kernel matching (where matches within a certain bandwidth are differentially weighted based on a kernel function). Imbens and Wooldgridge (2009) discuss the recent developments in matching methods.

Before calculating the effect of the intervention, it is important to examine the quality of the matches. This is typically accomplished by examine the standardized differences in the variables used in the analysis between the treatment and comparison groups. One rule of thumb is that standardized differences greater than 10 are problematic, and indicate that there may be a need to reevaluate the propensity score estimation method and the matching method.

The next step is to estimate the differences in outcomes between the matched treatment and comparison groups. Most statistical software packages can implement a number of methods for estimating these differences.

B.6 DIFFERENCE-IN-DIFFERENCE

To recover the difference-in-difference estimate of the program, data should be collected before and after the program for both the treatment and the comparison group. Then, the following regression equation can be estimated:

$$Y_i = \alpha_0 + \alpha_1 P_i + \alpha_0 + \alpha_2 T_i + \beta_{DID} T_i P_i + f(\boldsymbol{X}_i) + \varepsilon$$

where P_i is an indicator for after the financial capability intervention, and β_{DID} gives the estimate of the intervention effect. It is advisable to include controls for any additional characteristics of the treatment and comparison group that were collected. Here we have included them as $f(\boldsymbol{X}_i)$ to emphasize that flexible functions of these controls can be specified.

B.7 POWER CALCULATION

We are interested in testing the hypothesis that the effect of the program is zero. When you are designing an evaluation, it is important to determine the sample size required to estimate a particular effect size, or to analyze the power of a significance test for a given sample size. The power of a design is the probability that, for a given

effect size and a given statistical significance level we will be able to reject the hypothesis of zero effect.

In order to calculate the required sample size or determine the power of the design, a number of parameters need to be specified. α is the significance level of the test, and it represents the probability that we reject the hypothesis when it is in fact true. In general, α is usually assumed to be 0.05. Let *P* represent the proportion of the sample assigned to the treatment (usually assumed to equal 0.5), and *T* indicate treatment status.

The minimum detectible effect size for a randomized control trial for a given power ($1-\kappa$), significance level (α), sample size (*N*), and proportion allocated to the treatment group (*P*) is

$$MDE = (t_{(1-\kappa)} + t_{\alpha}) * \sqrt{\frac{1}{P(1-P)}} \sqrt{\frac{\sigma^2}{N}}$$

where t_{α} and $t_{(1-\kappa)}$ are taken from the standard *t*-distribution table, and σ^2 is the variance of the measurement error.

B.8 REGRESSION DISCONTINUITY DESIGN

Because in a regression discontinuity design only observations near the cutoff point are used in the estimation, the analysis requires additional power. According to Schochet (2009), the RD design effect depends on the extent to which the distribution of the assignment variable is truncated. Generally speaking, the more truncated the distribution, the greater the constraint on statistical power. Truncation, in turn, is determined by the bandwidth from the cross-validation technique cited above and depends on (1) the distribution of the assignment scores in the whole population, (2) the location of the threshold score in this distribution, and (3) the portion of students in the sample assigned to treatment or the comparison group.

The sample size for a given power ($1-\kappa$), significance level (α), minimum detectible effect size (*M*), and proportion allocated to the treatment group (*P*) is

$$N = \frac{(1-R_M^2)(t_{(1-\kappa)} + t_{\alpha})^2}{M^2 P(1-P)(1-R_T^2)}$$

where R_M^2 is the R-squared statistic for the OLS regression, and R_T^2 is the correlation between the assignment variable (*R*) and the treatment status (*T*). Lee and Lemieux (2009) provide a table of sample sizes for a range of minimum detectable effect size and R-squared statistic for the OLS regression assuming an α of 0.05, power of 0.8, proportion assigned to treatment of 0.5, and a correlation between the assignment variable of 0.667.

FURTHER READING

General

Imbens, G. W., and J. M. Wooldridge. 2009. "Recent Developments in the Econometrics of Program Evaluation." *Journal of Economic Literature* 47 (1), 5–86.

Wooldridge, J. M. 2002. *Econometric Analysis of Cross Section and Panel Data.* Cambridge, MA: MIT Press.

Randomized control trials

Duflo, E., R. Glennerster, and M. Kremer. 2007. "Using randomization in development economics research: A toolkit." *Handbook of Development Economics* 4: 3895–962.

Encouragement design and instrumental variables

Bradlow, E. T. 1998. "Encouragement Designs: An Approach to Self-Selected Samples in an Experimental Design." *Marketing Letters* 9 (4), 383–91.

Imbens, G. W., and J. D. Angrist. 1994. "Identification and Estimation of Local Average Treatment Effects." *Econometrica* 62 (2): 467–75.

Angrist, J. D., G. W. Imbens, and D. B. Rubin. 1996. "Identification of Causal Effects Using Instrumental Variables." *Journal of the American Statistical Association* 91 (434): 444–55.

Regression discontinuity

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

Lee, D. S., and T. Lemieux. 2010. "Regression Discontinuity Designs in Economics." *Journal of Economic Literature* 48 (2): 281–355.

Van der Klaauw, W. 2008. "Regression-Discontinuity Analysis: A Survey of Recent Developments in Economics." *Labour* 22 (2): 219–45.

Propensity score matching

Rosenbaum, P. R., and D. B. Rubin. 1983. "The Central Role of the Propensity Score in Observational Studies for Causal Effects." *Biometrika* 70 (1): 41–55.

Heckman, J.J., Ichimura, H. and Todd, P.E. (1998), "Matching as an Econometric Evaluation Estimator," *Review of Economic Studies*, 65, 261–294.

Austin, P. C. 2008. "A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003." *Statistics in Medicine* 27 (12): 2037–049.

Difference-in-difference

Angrist, Joshua D., and Alan B. Krueger. 1999. "Empirical Strategies in Labor Economics." In *Handbook of Labor Economics, Volume 3A*, ed. Orley Ashenfelter and David Card, 1277–366. New York: Elsevier Science.

Angrist, Joshua D., and Jorn-Steffen Pischke. 2009. *Mostly Harmless Econometrics: An Empiricist's Companion.* Princeton: Princeton University Press.

Power calculations

Aberson, C. L. 2010. *Applied Power Analysis for the Behavioral Science.*

Ellis, Paul D. 2010. *The Essential Guide to Effect Sizes: An Introduction to Statistical Power, Meta-Analysis and the Interpretation of Research Results.* Cambridge, UK: Cambridge University Press.

Schochet, P. Z. 2009. "Statistical power for regression discontinuity designs in education evaluations." *Journal of Educational and Behavioral Statistics* 34, 238–66.

The Russia Financial Literacy and Education Trust Fund was established in 2008 at the World Bank with funding provided by the Ministry of Finance of the Russian Federation. The work supported by the Trust Fund is jointly managed by the World Bank and the Organisation for Economic Co-operation and Development (OECD) and is directed toward improving public policies and programs to enhance financial knowledge and capabilities in low- and middle-income countries. This effort has focused on the review of national strategies for financial education, the development of methods for the measurement of financial knowledge and capabilities, methods for evaluating the impact and outcome of programs, and research applying these methods to programs in developing countries. The products of this program of work can be found at the Trust Fund website at:

www.finlitedu.org

THE WORLD BANK

MINISTRY OF FINANCE OF THE RUSSIAN FEDERATION

OECD