Evaluation of International Development Interventions

An Overview of Approaches and Methods

Jos Vaessen

Sebastian Lemire

Barbara Befani

IEG
Independent Evaluation Group

WORLD BANK GROUP
World Bank • IFC • MIGA

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

ATTRIBUTION

Please cite the report as: Vaessen, Jos, Sebastian Lemire, and Barbara Befani. 2020. *Evaluation of International Development Interventions: An Overview of Approaches and Methods.* Independent Evaluation Group. Washington, DC: World Bank.

MANAGING EDITORS
Jos Vaessen
Ariya Hagh

EDITING AND PRODUCTION
Amanda O'Brien

GRAPHIC DESIGN
Luísa Ulhoa

This work is a product of the staff of The World Bank with external contributions. The findings, interpretations, and conclusions expressed in this work do not necessarily reflect the views of The World Bank, its Board of Executive Directors, or the governments they represent.
The World Bank does not guarantee the accuracy of the data included in this work. The boundaries, colors, denominations, and other information shown on any map in this work do not imply any judgment on the part of The World Bank concerning the legal status of any territory or the endorsement or acceptance of such boundaries.

RIGHTS AND PERMISSIONS

Evaluation of International Development Interventions

An Overview of Approaches and Methods

Jos Vaessen, Sebastian Lemire, and Barbara Befani

Independent Evaluation Group

November 2020

CONTENTS

ABBREVIATIONS

IEO	independent evaluation office
QCA	qualitative comparative analysis
SMS	short message service
SNA	social network analysis
USAID	United States Agency for International Development

ACKNOWLEDGMENTS

The authors are grateful for the comments provided by the external peer reviewers: Ole Winckler Andersen, Pablo Guerrero, and Michael Bamberger. In addition, the authors would like to thank the following colleagues at the Independent Evaluation Group for their comments and contributions: Alison Evans, Ariya Hagh, Caroline Heider, Giuseppe Iarossi, Kristen Milhollin, Amanda O’Brien, Estelle Raimondo, Luísa Ulhoa, and Maurya West Meiers.

FOREWORD

In recent years, evaluation in the field of international development has undergone significant changes. First and foremost, the world we live in has become increasingly complex and interconnected. In the current era of the Sustainable Development Goals, governments, international organizations, private corporations, civil society organizations, and others are increasingly aware of the challenges surrounding transboundary and global issues such as climate change, migration, and international trade. At the same time, evaluation as a source of independent inquiry into the merit and worth of policy interventions to address global, national, and local challenges has grown in importance. The increased number of evaluation functions in governmental, nongovernmental, and private sector organizations; the growing number of countries with voluntary professional organizations for evaluators; and the growth in repositories of knowledge on policy interventions and their (potential) effects are all signs of this trend.

How should evaluators deal with the increasing complexity of policy interventions and the contexts in which they potentially influence change? How can evaluation as a function, as a practice, effectively contribute to the evolving learning and accountability needs of decision makers, practitioners, financiers, and citizens? These are questions that have no easy answers and to some extent require a departure from how things were done in the past. For example, interventions that influence the lives of the poor, the distribution of wealth, or the sustainable use and conservation of natural resources are often strongly interconnected. Ideally, in such cases these interventions should not be assessed in isolation from each other. Similarly, decision makers and other stakeholders no longer rely on activity-level or project-level evaluations only. Assessments of programs, strategies, or even a comprehensive range of interventions that have a bearing on the same phenomenon are becoming more important as “evaluands.” Multisectoral, multidimensional, and multistakeholder perspectives on change and the way policy interventions affect change are called for. Finally, new technologies for data collection and analysis (for example, machine learning) and new types of data (for example, “big data”) are slowly but steadily making their entry into the practice of evaluation.

In light of these challenges, evaluators should broaden their methodological repertoire so that they are better able to match the evaluation questions and the operational constraints of the evaluation to the right methodological approach. Eminent development thinkers and evaluation scholars, such as Albert

Hirschman, Peter Rossi, and Ray Pawson, have called evaluation applied social science research. Evaluators should look to the social sciences when developing their methodological designs and realize that even within the boundaries of their mandates and institutions, there are many opportunities to develop, test, and apply modern methods of data collection and analysis. It is in doing so, and in combining a mix of methods that corresponds to the specificities of the evaluation at hand, that evaluators can provide new insights into development interventions and their consequences.

This guide provides an overview of evaluation approaches and methods that have been used in the field of international development evaluation. Although by no means exhaustive, the modules provide accessible and concise information on useful approaches and methods that have been selected for their actual use and their potential in evaluation. The reading lists at the end of each module refer the reader to useful resources to learn more about the applicability and utility of the methods described. Both the choice of approaches and methods and the associated guidance are by no means definitive. We hope to update this information as evaluation practices evolve.

We hope that this resource will be helpful to evaluators and other evaluation stakeholders alike and inform their practice.

Alison Evans
Director-General, Evaluation

1

GUIDANCE TO THE READER

Scope and Aim

Intended Audience

Selected Approaches and Methods

Structure of the Guide

Scope and Aim

This guide is intended as a quick reference to evaluation approaches and methods.[1] Its aim is to provide easily accessible and jargon-light descriptions of a broad yet select set of methodological approaches and methods used in evaluations in the field of international development (and beyond). The guide provides concise descriptions of established and emerging approaches and methods. The featured selections were chosen especially for readers interested in independent evaluation in international development. At the same time, we think that the approaches and methods reflected in the guide will be relevant to a much broader audience of evaluators and policy researchers.

The guide is inevitably selective in its coverage and by no means all-inclusive, reflecting what we consider some of the most salient current trends in development evaluation practice.[2, 3]

In our discussion of the selected methodological approaches, we have tried to keep the level of complexity and technical detail to a minimum, focusing on the following key aspects: a short description of the approach or method, main steps involved in its application, variations in methodological principles or application, advantages and disadvantages, and applicability. In addition, we provide some examples of applications of the approach or method and references to the literature (both basic and more advanced). Consequently, the guide will help evaluation stakeholders to become more aware of different methods and approaches and gain practical insights regarding their applicability and where to look for additional guidance.

The guide is *not* intended as a step-by-step manual on how to design and conduct evaluations. This type of guidance is already provided in a number of widely used publications (see, for example, Bamberger, Rugh, and Mabry 2006; Morra Imas and Rist 2009). Similarly, we will not discuss the ontological and epistemological foundations of the included approaches and methods.[4] These debates, although interesting in their own right, have been covered well in other publications (see, for example, Pawson and Tilley 1997; Alkin 2004; Stern et al. 2012, among others) and are outside the scope of this guide. In the end, and despite these boundaries, our modest hope is that the guide will broaden readers' methodological knowledge and inform their future design, conduct, and use of evaluations.

Finally, a central message of the guide is that there is no single "best" evaluation approach or method. The approach should be determined by the nature of the intervention being evaluated, the types of questions the evaluation addresses, and the opportunities and constraints under which the evaluation is conducted in terms of available time, data, budget, and institutional constraints and preferences.

The Intended Audience of the Guide

We expect that a variety of professionals who are involved with evaluation in some way would find this guide useful: novices entering the evaluation field; experienced evaluators interested in quick-access summaries of a range of established and emerging approaches and methods; project managers or commissioners of evaluations who might not necessarily have a background in evaluation methods but are nevertheless involved in the evaluation function; and so on. Similarly, most of the approaches and methods would be of interest to evaluation stakeholders in a range of institutional settings: multilateral or bilateral organizations, government agencies, nongovernmental organizations, private sector organizations, academia, and other bodies. In addition, professionals working in program planning, management, or monitoring and related roles may also find the content useful. Finally, policy-oriented researchers in international development may also find this guide useful as a quick reference. There is, however, a clear and intentional bias toward the work of independent evaluation offices (IEOs) as found in many multilateral development organizations (for example, multilateral development banks, United Nations agencies and programs), bilateral organizations, international nongovernmental organizations, or foundations.[5]

The Selected Approaches and Methods in the Guide

Because this guide is explicitly biased toward IEOs, much attention is given to summative evaluation approaches and methods and relatively less attention to formative (including developmental) approaches and methods for evaluation.[6]

The mandate of most IEOs influences the kinds of evaluation methods they are likely to use. Most evaluations are conducted either after the intervention (for example, project, sector program, or policy) has been completed (retrospective or ex post evaluation) or during an ongoing program or portfolio of interventions. By definition, *independence* implies that the evaluators are not directly involved in the design or implementation of the organization’s projects, programs, or policies. Furthermore, independence often requires that the IEO has little control over the operational arm of the organization and the kinds of information (useful to retrospective evaluation) that are collected during project design or implementation. Finally, IEO evaluations often operate at higher levels of analysis (for example, country or regional programs, thematic strategies), which influence the extent to which participatory methods can be (comprehensively) applied. Also, there is often a trade-off between breadth and depth of analysis that influences evaluation design and the scope for in-depth (causal) analysis. For these reasons, a number of approaches and methods—some of which are included in this guide (for example, experimental designs)—are often less suited and less commonly applied in IEO evaluations.

Recognizing these methodological “quasi-boundaries,” the guide mainly focuses on approaches and methods that can be used in retrospective (ex post) evaluations. At the same time, although many IEOs cannot regularly use some of the evaluation approaches and methods described in the guide, there are many exceptions. The increasing diversity in evaluation modalities and levels of evaluation (for example, global strategy, country program, thematic area of work, project) that IEOs are engaged in requires the application of a broader range of evaluation approaches.

It should be noted that this guide is intended to be a living document. As new relevant methods and applications for them emerge, we aim to periodically update the guidance notes.

The Structure of the Guide

The remainder of the guide is structured in two chapters. In chapter 2, Methodological Principles of Evaluation Design, we discuss seven guiding principles for designing quality evaluations in a development context, emphasizing the importance of allowing evaluation questions to drive methodological decisions, building program theory on stakeholder and substantive theory, mixing methods and approaches, balancing scope and depth, attending to context, and adapting approaches and methods to real-world constraints. The section is not intended as any sort of comprehensive guide to evaluation design. Rather, it examines methods choice according to a number of core methodological principles that evaluators may wish to reflect on.

Chapter 3, Guidance Notes on Evaluation Approaches and Methods in Development, presents an overview of select methodological approaches and more specific methods and tools. Each guidance note briefly describes the approach and its main variations, procedural steps, advantages and disadvantages, and applicability. Case examples and additional references and resources for each approach are provided.

References

Alkin, M. 2004. *Evaluation Roots: A Wider Perspective of Theorists’ Views and Influences*. Thousand Oaks, CA: SAGE.

Bamberger, M., J. Rugh, and L. Mabry. 2006. *RealWorld Evaluation: Working under Budget, Time, Data, and Political Constraints*. Thousand Oaks, CA: SAGE.

Morra Imas, L., and R. Rist. 2009. *The Road to Results*. Washington, DC: World Bank.

Pawson, R., and N. Tilley. 1997. *Realistic Evaluation*. Thousand Oaks, CA: SAGE.

Stern, E., N. Stame, J. Mayne, K. Forss, R. Davies, and B. Befani. 2012. “Broadening the Range of Designs and Methods for Impact Evaluations.” Working Paper 38, Department for International Development, London. https://www.oecd.org/derec/50399683.pdf.

Endnotes

[1] For simplification purposes we define *method* as a particular technique involving a set of principles to collect or analyze data, or both. The term *approach* can be situated at a more aggregate level, that is, at the level of methodology, and usually involves a combination of methods within a unified framework. Methodology provides the structure and principles for developing and supporting a particular knowledge claim.

[2] Development evaluation is not to be confused with developmental evaluation. The latter is a specific evaluation approach developed by Michael Patton.

[3] Especially in independent evaluations conducted by independent evaluation units or departments in national or international nongovernmental, governmental, and multilateral organizations. Although a broader range of evaluation approaches may be relevant to the practice of development evaluation, we consider the current selection to be at the core of evaluative practice in independent evaluation.

[4] From a philosophy of science perspective, the terms *ontology* and *epistemology*, respectively, refer to "how one views the world" and "what knowledge is."

[5] Evaluation functions of organizations that are (to a large extent), structurally, organizationally and behaviorally independent from management. Structural independence, which is the most distinguishing feature of independent evaluation offices, includes such aspects as independent budgets, independent human resource management, and no reporting line to management, but some type of oversight body (for example, an executive board).

[6] The latter are not fully excluded from this guide but are not widely covered.

2

METHODOLOGICAL PRINCIPLES OF EVALUATION DESIGN

Giving due consideration to methodological aspects of evaluation quality in design

Using effective tools for evaluation design

Mixing methods for analytical depth and breadth

Building on theory

Matching evaluation design to the evaluation questions

Balancing scope and depth in multilevel, multisite evaluands

Dealing with institutional opportunities and constraints of budget, data, and time

Evaluation approaches and methods do not exist in a vacuum. Stakeholders who commission or use evaluations and those who manage or conduct evaluations all have their own ideas and preferences about which approaches and methods to use. An individual's disciplinary background, experience, and institutional role influence such preferences; other factors include internalized ideas about rigor and applicability of methods. This guide will inform evaluation stakeholders about a range of approaches and methods that are used in evaluative analysis and provide a quick overview of the key features of each. It thus will inform them about the approaches and methods that work best in given situations.

Before we present the specific approaches and methods in chapter 3, let us consider some of the key methodological principles of evaluation design that provide the foundations for the selection, adaptation, and use of evaluation approaches and methods in an IEO evaluation setting. To be clear, we focus only on methodological issues here and do not discuss other key aspects of design, such as particular stakeholders' intended use of the evaluation. The principles discussed in this chapter pertain also to evaluation in general, but they are especially pertinent for designing independent evaluations in an international development context. We consider the following methodological principles to be important for developing high-quality evaluations:

1. Giving due consideration to methodological aspects of evaluation quality in design: focus, consistency, reliability, and validity
2. Matching evaluation design to the evaluation questions
3. Using effective tools for evaluation design
4. Balancing scope and depth in multilevel, multisite evaluands
5. Mixing methods for analytical depth and breadth
6. Dealing with institutional opportunities and constraints of budget, data, and time
7. Building on theory

Let us briefly review each of these in turn.

Giving Due Consideration to Methodological Aspects of Evaluation Quality in Design

Evaluation quality is complex. It may be interpreted in different ways and refer to one or more aspects of quality in terms of process, use of methods, team composition, findings, and so on. Here we will talk about quality of inference: the quality of the findings of an evaluation as underpinned by clear reasoning and reliable evidence. We can differentiate among four broad, interrelated sets of determinants:

- The budget, data, and time available for an evaluation (see the Dealing with Institutional Opportunities and Constraints of Budget, Data, and Time section);
- The institutional processes and incentives for producing quality work;
- The expertise available within the evaluation team in terms of different types of knowledge and experience relevant to the evaluation: institutional, subject matter, contextual (for example, country), methodological, project management, communication; and
- Overarching principles of quality of inference in evaluation research based on our experience and the methodological literature in the social and behavioral sciences.[1]

Here we briefly discuss the final bullet point. From a methodological perspective, quality can be broken down into four aspects: focus, consistency, reliability, and validity.

Focus concerns the scope of the evaluation. Given the nature of the evaluand and the type of questions, how narrowly or widely does one cast the net? Does one look at both relevance and effectiveness issues? How far down the causal chain does the evaluation try to capture the causal contribution of an intervention? Essentially, the narrower the focus of an evaluation, the greater the concentration of financial and human resources on a particular aspect and consequently the greater the likelihood of high-quality inference.

Consistency here refers to the extent to which the different analytical steps of an evaluation are logically connected. The quality of inference is enhanced if there are logical connections among the initial problem statement, rationale and purpose of the evaluation, questions and scope, use of methods, data collection and analysis, and conclusions of an evaluation.

Reliability concerns the transparency and replicability of the evaluation process.[2] The more systematic the evaluation process and the higher the levels of clarity and transparency of design and implementation, the higher the confidence of others in the quality of inference.

Finally, validity is a property of findings. There are many classifications of validity. A widely used typology is the one developed by Cook and Campbell (1979) and slightly refined by Hedges (2017):

- Internal validity: To what extent is there a causal relationship between, for example, outputs and outcomes?

- External validity: To what extent can we generalize findings to other contexts, people, or time periods?
- Construct validity: To what extent is the element that we have measured a good representation of the phenomenon we are interested in?
- Data analysis validity: To what extent are methods applied correctly and the data used in the analysis adequate for drawing conclusions?

Matching Evaluation Design to the Evaluation Questions

Although it may seem obvious that evaluation design should be matched to the evaluation questions, in practice much evaluation design is still too often methods driven. Evaluation professionals have implicit and explicit preferences and biases toward the approaches and methods they favor. The rise in randomized experiments for causal analysis is largely the result of a methods-driven movement. Although this guide is not the place to discuss whether methods-driven evaluation is justified, there are strong arguments against it. One such argument is that in IEOs (and in many similar institutional settings), one does not have the luxury of being too methods driven. In fact, the evaluation questions, types of evaluands, or types of outcomes that decision makers or other evaluation stakeholders are interested in are diverse and do not lend themselves to one singular approach or method for evaluation. Even for a subset of causal questions, given the nature of the evaluands and outcomes of interest (for example, the effect of technical assistance on institutional reform versus the effect of micrograns on health-seeking behavior of poor women), the availability and cost of data, and many other factors, there is never one single approach or method that is always better than others. For particular types of questions there are usually several methodological options with different requirements and characteristics that are better suited than others. Multiple classifications of questions can be helpful to evaluators in thinking more systematically about this link, such as causal versus noncausal questions, descriptive versus analytical questions, normative versus nonnormative questions, intervention-focused versus systems-based questions, and so on. Throughout this guide, each guidance note presents what we take to be the most relevant questions that the approach or method addresses.

Using Effective Tools for Evaluation Design

Over the years, the international evaluation community in general and institutionalized evaluation functions (such as IEOs) in particular have developed and used a number of tools to improve the quality and efficiency of evaluation design.[3] Let us briefly discuss four prevalent tools.

First, a common tool in IEOs (and similar evaluation functions) is some type of multicriteria approach to justify the strategic selectivity of topics or interventions for evaluation. This could include demand-driven criteria such as potential stakeholder use or supply-driven criteria such as the financial volume or size of a program or portfolio of interventions. Strategic selectivity often goes hand in hand with evaluability assessment (Wholey 1979), which covers such aspects as stakeholder interest and potential use, data availability, and clarity of the evaluand (for example, whether a clear theory of change underlies the evaluand).

A second important tool is the use of approach papers or inception reports. These are stand-alone documents that describe key considerations and decisions regarding the rationale, scope, and methodology of an evaluation. When evaluations are contracted out, the terms of reference for external consultants often contain similar elements. Terms of reference are, however, never a substitute for approach papers or inception reports.

As part of approach papers and inception reports, a third tool is the use of a design matrix. For each of the main evaluation questions, this matrix specifies the sources of evidence and the use of methods. Design matrixes may also be structured to reflect the multilevel nature (for example, global, selected countries, selected interventions) of the evaluation.

A fourth tool is the use of external peer reviewers or a reference group. Including external methodological and substantive experts in the evaluation design process can effectively reduce bias and enhance quality.

Balancing Scope and Depth in Multilevel, Multisite Evaluands

Although project-level evaluation continues to be important, at the same time and for multiple reasons international organizations and national governments are increasingly commissioning and conducting evaluations at higher programmatic levels of intervention. Examples of the latter are sector-level evaluations, country program evaluations, and regional or global thematic evaluations. These evaluations tend to have the following characteristics:

- They often cover multiple levels of intervention, multiple sites (communities, provinces, countries), and multiple stakeholder groups at different levels and sites.
- They are usually more summative and are useful for accountability purposes, but they may also contain important lessons for oversight bodies, management, operations, or other stakeholders.
- They are characterized by elaborate evaluation designs.

A number of key considerations for evaluation design are specific to higher-level programmatic evaluations. The multilevel nature of the intervention (portfolio) requires a multilevel design with multiple methods applied at different levels of analysis (such as country or intervention type). For example, a national program to support the health sector in a given country may have interventions relating to policy dialogue, policy advisory support, and technical capacity development at the level of the line ministry while supporting particular health system and health service delivery activities across the country. Multilevel methods choice goes hand in hand with multilevel sampling and selection issues. A global evaluation of an international organization's support to private sector development may involve data collection and analysis at the global level (for example, global institutional mapping), the level of the organization's portfolio (for example, desk review), the level of selected countries (for example, interviews with representatives of selected government departments or agencies and industry leaders), and the level of selected interventions (for example, theory-based causal analysis of advisory services in the energy sector). For efficiency, designs are often "nested"; for example, the evaluation covers selected interventions in selected countries. Evaluation designs may encompass different case study levels, with within-case analysis in a specific country (or regarding a specific intervention) and cross-case (comparative) analysis across countries (or interventions). A key constraint in this type of evaluation is that one cannot cover everything. Even for one evaluation question, decisions on selectivity and scope are needed. Consequently, strategic questions should address the desired breadth and depth of analysis. In general, the need for depth of analysis (determined by, for example, the time, resources, and triangulation among methods needed to understand and assess one particular phenomenon) must be balanced by the need to generate generalizable claims (through informed sampling and selection). In addition to informed sampling and selection, generalizability of findings is influenced by the degree of convergence of findings from one or more cases with available existing evidence or of findings across cases. In addition, there is a clear need for breadth of analysis in an evaluation (looking at multiple questions, phenomena, and underlying factors) to adequately cover the scope of the evaluation. All these considerations require careful reflection in what can be a quite complicated evaluation design process.

Mixing Methods for Analytical Depth and Breadth

Multilevel, multisite evaluations are by definition multimethod evaluations. But the idea of informed evaluation design, or the strategic mixing of methods applies to essentially all evaluations. According to Bamberger (2012, 1), "Mixed methods evaluations seek to integrate social science disciplines with predominantly quantitative and predominantly qualitative approaches to theory, data collection, data analysis and interpretation. The purpose is to strengthen the reliability of data, validity of the findings and recommendations, and to broaden and deepen our understanding of the processes through which program outcomes and impacts are achieved, and how these are affected by the context within which the program is implemented.” The evaluator should always strive to identify and use the best-suited methods for the specific purposes and context of the evaluation and consider how other methods may compensate for any limitations of the selected methods. Although it is difficult to truly integrate different methods within a single evaluation design, the benefits of mixed methods designs are worth pursuing in most situations. The benefits are not just methodological; through mixed designs and methods, evaluations are better able to answer a broader range of questions and more aspects of each question.

There is an extensive and growing literature on mixed methods in evaluation. One of the seminal articles on the subject (by Greene, Caracelli, and Graham) provides a clear framework for using mixed methods in evaluation that is as relevant as ever. Greene, Caracelli, and Graham (1989) identify the following five principles and purposes of mixing methods:

Triangulation Using different methods to compare findings. Convergence of findings from multiple methods strengthens the validity of findings. For example, a survey on investment behavior administered to a random sample of owners of small enterprises could confirm the findings obtained from semi-structured interviews for a purposive sample of representatives of investment companies supporting the enterprises.

Initiation Using different methods to critically question a particular position or line of thought. For example, an evaluator could test two rival theories (with different underlying methods) on the causal relationships between promoting alternative livelihoods in buffer zones of protected areas and protecting biodiversity.

Complementarity Using one method to build on the findings from another method. For example, in-depth interviews with selected households and their individual members could deepen the findings from a quasi-experimental analysis on the relationship between advocacy campaigns and health-seeking behavior.

Development Using one method to inform the development of another. For example, focus groups could be used to develop a contextualized understanding of women's empowerment and could use that information to develop a survey questionnaire.

Expansion Using multiple methods to look at complementary areas. For example, social network analysis could be used to understand an organization's position in the financial landscape of all major organizations supporting a country's education sector, while using semistructured interviews with officials from the education ministry and related agencies to assess the relevance of the organization's support to the sector.

Dealing with Institutional Opportunities and Constraints of Budget, Data, and Time

Evaluation is applied social science research in the context of specific institutional requirements, constraints, and opportunities, and a range of other practical constraints. Addressing these all-too-common constraints, including budget, data, time, political, and other constraints, involves balancing rigor and depth of analysis with feasibility. In this sense, evaluation clearly distinguishes itself from academic research in several ways:

- It is strongly linked to an organization's accountability and learning processes, and there is some explicit or implicit demand-orientation in evaluation.
- It is highly normative, and evidence is used to underpin normative conclusions about the merit and worth of an evaluand.
- It puts the policy intervention (for example, the program, strategy, project, corporate process, thematic area of work) at the center of the analysis.
- It is subject to institutional constraints of budget, time, and data. Even in more complicated evaluations of larger programmatic evaluands, evaluation (especially by IEOs) is essentially about "finding out fast" without compromising too much the quality of the analysis.
- It is shaped in part by the availability of data already in the organizational system. Such data may include corporate data (financial, human resources, procurement, and so on), existing reporting (financial appraisal, monitoring, [self-] evaluation), and other data and background research conducted by the organization or its partners.

Building on Theory

Interventions are theories, and evaluation is the test (Pawson and Tilley 2001). This well-known reference indicates an influential school of thought and practice in evaluation, often called theory-driven or theory-based evaluation. Policy interventions (programs and projects) rely on underlying theories regarding how they are intended to work and contribute to processes of change. These theories (usually called program theories, theories of change, or intervention theories) are often made explicit in documents but sometimes exist only in the minds of stakeholders (for example, decision makers, evaluation commissioners, implementing staff, beneficiaries). Program theories (whether explicit or tacit) guide the design and implementation of policy interventions and also constitute an important basis for evaluation.

The important role of program theory (or variants thereof) is well established in evaluation. By describing the inner workings of how programs operate (or at least are intended to operate), the use of program theory is a fundamental step in evaluation planning and design. Regardless of the evaluation question or purpose, a central step will always be to develop a thorough understanding of the intervention that is evaluated. To this end, the development of program theories should always be grounded in stakeholder knowledge and informed to the extent possible by social scientific theories from psychology, sociology, economics, and other disciplines. Building program theories on the basis of stakeholder knowledge and social scientific theory supports more relevant and practice-grounded program theories, improves the conceptual clarity and precision of the theories, and ultimately increases the credibility of the evaluation.

Depending on the level of complexity of the evaluand (for example, a complex global portfolio on urban infrastructure support versus a specific road construction project) a program theory can serve as an overall sense-making framework; a framework for evaluation design by linking particular causal steps and assumptions to methods and data; or a framework for systematic causal analysis (for example, using qualitative comparative analysis or process tracing; see chapter 3). Program theories can be nested; more detailed theories of selected (sets of) interventions can be developed and used for guiding data collection, analysis, and the interpretation of findings, while the broader theory can be used to connect the different strands of intervention activities and to make sense of the broader evaluand (see also appendix B).

References

Bamberger, M. 2012. *Introduction to Mixed Methods in Impact Evaluation*. Impact Evaluation Notes 3 (August), InterAction and the Rockefeller Foundation. https://www.interaction.org/wp-content/uploads/2019/03/Mixed-Methods-in-Impact-Evaluation-English.pdf.

Bamberger, M., J. Rugh, and L. Mabry. 2006. *RealWorld Evaluation: Working under Budget, Time, Data, and Political Constraints*. Thousand Oaks, CA: SAGE.

Cook, T. D., and D. T. Campbell. 1979. *Quasi-Experimentation: Design and Analysis Issues for Field Settings*. Boston: Houghton Mifflin.

Greene, J., V. Caracelli, and W. Graham. 1989. "Toward a Conceptual Framework for Mixed-Method Evaluation Designs." *Educational Evaluation and Policy Analysis* 11 (3): 209–21.

Hedges, L. V. 2017. "Design of Empirical Research." In *Research Methods and Methodologies in Education*, 2nd ed., edited by R. Coe, M. Waring, L. V. Hedges, and J. Arthur, 25–33. Thousand Oaks, CA: SAGE.

Morra Imas, L., and R. Rist. 2009. *The Road to Results*. Washington, DC: World Bank.

Pawson, R., and N. Tilley. 2001. "Realistic Evaluation Bloodlines." *American Journal of Evaluation* 22 (3): 317–24.

Wholey, Joseph. 1979. *Evaluation—Promise and Performance*. Washington, DC: Urban Institute.

Endnotes

[1] Evaluation is defined as applied policy-oriented research and builds on the principles, theories, and methods of the social and behavioral sciences.

[2] Both reliability and validity are covered by a broad literature. Many of the ideas about these two principles are contested, and perspectives differ according to different schools of thought (with different underlying ontological and epistemological foundations).

[3] A comprehensive discussion of the evaluation process, including tools, processes, and standards for designing, managing, quality assuring, disseminating, and using evaluations is effectively outside of the scope of this guide (see instead, for example, Bamberger, Rugh, and Mabry 2006; Morra Imas and Rist 2009).

3

GUIDANCE NOTES ON EVALUATION APPROACHES AND METHODS IN DEVELOPMENT

This chapter presents guidance notes on approaches and methods for evaluators working in international development. First, we describe several prevalent methodological approaches, followed by guidance notes on specific methods and tools. The latter are further divided according to their primary function for either *data collection* or *data analysis*.

Each guidance note briefly describes the approach and its methodological variations, main procedural steps, key advantages and disadvantages, and applicability. Finally, each guidance note includes references for relevant background readings (for example, journal articles, book chapters, method guides), illustrative case examples, and, when relevant, other useful resources (for example, links to useful online tools or software).

MAIN METHODOLOGICAL APPROACHES

1 Efficiency Analysis: Cost-Benefit and Cost-Effectiveness

BRIEF DESCRIPTION OF THE APPROACH

Efficiency analysis commonly refers to economic approaches that compare the relative costs and benefits (outcomes) of the program being evaluated. The main reason to do efficiency analysis is to establish whether the benefits of the program outweigh its associated costs. This type of information is particularly relevant when making decisions about future program planning and design and when considering alternative programs. Evaluation questions to be answered by efficiency analysis include the following:

1. What is the (cumulative) program effect relative to program costs?
2. To what extent does the benefit-cost ratio of the program vary across subgroups of the population?
3. How does the cost-effectiveness of the program compare with that of other programs (or other program variants)?

THE MAIN VARIATIONS OF THE APPROACH

Two main variations of efficiency analysis are *cost-benefit* and *cost-effectiveness* analysis. In cost-benefit analysis, also known as benefit-cost analysis, the program costs and effects are both defined in monetary terms, allowing for a direct comparison of costs and effects. The analysis can be conducted from a strictly financial or more general economic perspective.

In contrast, cost-effectiveness analysis compares the program costs defined in monetary terms with program effects defined in nonmonetary terms. For example, the number of children vaccinated may be a program effect in cost-effective analysis. Moreover, cost-effectiveness analyses often involve the comparison of cost-effectiveness ratios of similar programs or program variations.

Other closely related variants include cost-utility analysis, risk-benefit analysis, and social return on investment analysis, among others.

THE MAIN PROCEDURAL STEPS OF THE APPROACH

Efficiency analysis usually involves seven core steps:

- Defining the program costs and effects (effects are usually specified by program objectives);
- Deciding which costs and benefits should be included;
- Estimating the program costs;
- Quantifying the net program benefits (in monetary terms for cost-benefit analysis);
- Adjusting costs and benefits to net present value using a discount rate (see *discount rate* in appendix A, Glossary of Key Terms);
- Calculating the estimated cost-effectiveness ratio (or net [present] value for cost-benefit analysis); and
- Conducting robustness checks and sensitivity analysis.

The cornerstone of any efficiency analysis is the careful and thorough identification and measurement of all cost elements conceivably related to the program being evaluated. Conceptually, costs are defined as the sum of all program resources: staffing, supplies, facilities, and so on. Although many of these costs can be measured in monetary terms and valued through program records, other costs, such as in-kind contributions or other indirect costs incurred by partnering agencies, can be more difficult to accurately identify and quantify. For cost-benefit analysis, quantifying and assigning monetary values to the benefits constitutes a second cornerstone.

THE ADVANTAGES AND DISADVANTAGES OF THE APPROACH

Efficiency analysis can be conducted before or after a program has been implemented (that is, prospectively or retrospectively). If designed and implemented well, findings from efficiency analyses may serve well to inform future program planning and designs, for example, by motivating scale-up of cost-effective programs and modifications of cost-ineffective programs. Establishing cost-effectiveness for accountability purposes may also justify incurred program costs and support continuation of funding, especially in the context of budget or resource constraints. But the results of efficiency analyses are only one of many different inputs on which these decisions are and should be made.

There are also significant challenges in the use of efficiency analyses. Quantifying program costs and benefits (and any negative effects) in monetary terms can be difficult, especially when defining and measuring outcomes such as resilience, empowerment, or safety. Even if program costs and effects can be conceptually and thoughtfully defined, collecting relevant data can also be difficult. Designing and implementing rigorous efficiency analyses demands high capacity of the team in terms of economic and financial analysis, statistics, and program knowledge. Finally, the quality of cost-benefit analyses can be difficult to assess when there is limited transparency on how costs and benefits are defined, identified, and measured. The development and inclusion of a table specifying the included program costs and the provision of clear specifications of the program effects is considered standard practice.

THE APPLICABILITY OF THE APPROACH

Efficiency analyses are most useful to those involved in the design and oversight of programs, and, importantly, to those whom the program would directly affect, positively or negatively. Efficiency analyses may prompt decision makers to consider alternative program options, including taking no action when costs outweigh benefits. Efficiency analyses can provide significant value to independent evaluators as inputs to their broader evaluation studies—if they are implemented competently and transparently.

Practical applications of cost-effectiveness and cost-benefit analysis include the following:

1. A retrospective social cost-benefit analysis was used for cholera vaccination in Sub-Saharan Africa. Based on economic and epidemiological data collected in Beira, Mozambique, the analysis compares the net economic benefits of three immunization strategies.

   (*Source*: Jeuland, M., M. Lucas, J. Clemens, and D. Whittington. 2009. "A Cost-Benefit Analysis of Cholera Vaccination Programs in Beira, Mozambique." *The World Bank Economic Review* 23 (2): 235–67. https://openknowledge.worldbank.org/handle/10986/4502.)

2. Cost-benefit analysis was used for the AGEXPORT, a rural value chains project in Guatemala, where local producer associations were supported financially to help their members produce high-quality coffee beans. Net benefits were broken down by small, medium-size, and large farms.

(*Source*: USAID [US Agency for International Development]. 2013. *Economic Analysis of Feed the Future Investments—Guatemala*. Washington, DC: USAID. https://www.usaid.gov/documents/1865/economic-analysis-feed-future-investments-guatemala.)

3. Cost-benefit analysis was used for the Markets II program in Nigeria, where multiple interventions for poor rural farmers sought to improve their access to better inputs, adequate finance, better water management, appropriate technology, extension services, and improved nutritional uses of grown or purchased basic foods. The results at the farm level were aggregated and projected for a 10-year prognosis.

   (*Source*: USAID [US Agency for International Development]. 2015. *Cost-Benefit Analysis of USAID/Nigeria's Markets II Program*. Washington, DC: USAID. https://www.usaid.gov/documents/1865/cost-benefit-analysis-usaidnigeria%E2%80%99s-markets-ii-program#overlay-context=what-we-do/economic-growth-and-trade/promoting-sound-economic-policies-growth/working-more.)

4. Cost-effectiveness analysis was used to compare multiple education interventions for increasing school attendance in Kenya. Each intervention was subjected to a randomized design and cost-benefit analysis.

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

5. An ex post cost-benefit analysis was used to assess the economic justification of 50 World Bank–financed dam projects.

   (*Source*: World Bank. 2005. *Influential Evaluations: Detailed Case Studies*. Washington, DC: World Bank. http://documents.worldbank.org/curated/en/928001468330038416/pdf/328800Influent1luation1case1studies.pdf.)

6. A prospective cost-benefit analysis was applied on an earthquake vulnerability reduction project in Colombia.

   (*Source*: Ghesquiere, F., L. Jamin, and O. Mahul. 2006. "Earthquake Vulnerability Reduction Program in Colombia: A Probabilistic Cost-Benefit Analysis." Policy Research Working Paper 3939, World

Bank, Washington, DC. https://openknowledge.worldbank.org/handle/10986/8438.)

7. Cost-benefit analysis was used to assess land fragmentation and its impact on the efficiency of resource use in rural Rwanda.

   (*Source*: Ali, D. A., K. Deininger, and L. Ronchi. 2015. "Costs and Benefits of Land Fragmentation: Evidence from Rwanda." Policy Research Working Paper 7290, World Bank, Washington, DC. https://openknowledge.worldbank.org/handle/10986/22163.)

8. A retroactive cost-benefit analysis was used in a reassessment of the welfare impact of rural electrification programs.

   (*Source*: World Bank. 2008. *The Welfare Impact of Rural Electrification: A Reassessment of the Costs and Benefits*. Washington, DC: World Bank. https://openknowledge.worldbank.org/handle/10986/6519.)

9. A retrospective cost-benefit analysis of six road projects was conducted in Argentina, Botswana, India, Kenya, the Lao People's Democratic Republic, and Paraguay. The findings were used for scenario analysis to project economic viability of future road projects.

   (*Source*: Tsunokawa, K. 2010. *Road Projects Cost Benefit Analysis: Scenario Analysis of the Effect of Varying Inputs*. Working Paper 81577, World Bank, Washington, DC. https://openknowledge.worldbank.org/handle/10986/27814.)

READINGS AND RESOURCES

Background

Asian Development Bank. 2013. *Cost-Benefit Analysis for Development: A Practical Guide*. Mandaluyong City, Philippines: Asian Development Bank. https://www.adb.org/documents/cost-benefit-analysis-development-practical-guide.

Belli, P., J. R. Anderson, H. N. Barnum, J. A. Dixon, and J-. P. Tan. 2001. *Economic Analysis of Investment Operations—Analytical Tools and Practical Applications*. Washington, DC: World Bank. http://documents.worldbank.org/curated/en/792771468323717830/pdf/298210REPLACEMENT.pdf.

Boardman, A., A. Vining, D. Greenberg, and D. Weimer. 2001. *Cost-Benefit Analysis: Concepts and Practice*. New Jersey, NJ: Prentice Hall.

Dhaliwal, Iqbal, Esther Duflo, Rachel Glennerster, and Caitlin Tulloch. 2011. "Comparative Cost-Effectiveness Analysis to Inform Policy in Developing Countries: A General Framework with Applications for Education." Abdul Latif Jameel Poverty Action Lab (J-PAL), MIT. Cambridge, MA: Massachusetts Institute of Technology. https://economics.mit.edu/files/6959/.

Fletcher, J. D. 2010. "Cost Analysis in Evaluation Studies." In *International Encyclopedia of Education*, edited by Penelope Peterson, Eva Baker, and Barry McGaw, 585–91. Oxford: Elsevier.

Jamison, Dean T. 2009. "Cost Effectiveness Analysis: Concepts and Applications." In *Methods of Public Health*, vol. 2 of *Oxford Textbook of Public Health*, 5th ed., edited by R. Detels, J. McEwen, R. Beaglehole, and H. Tanaka, 767–82. Oxford: Oxford University Press. http://depts.washington.edu/cfar/sites/default/files/uploads/core-program/user164/Jamison%20CEA%20Concepts%20and%20Applications.pdf.

Levin, Henry M., Patrick J. McEwan, Clive R. Belfield, A. Brooks Bowden, and Robert D. Shand. 2018. *Economic Evaluation in Education: Cost-Effectiveness and Benefit-Cost Analysis*, 3rd ed. Thousand Oaks, CA: SAGE.

Little, I. M. D., and James A. Mirrlees. 1974. *Project Appraisal and Planning for Developing Countries*. London: Heinemann Educational Books.

McEwan, Patrick J. 2012. "Cost-Effectiveness Analysis of Education and Health Interventions in Developing Countries." *Journal of Development Effectiveness* 4 (2): 189–213. http://academics.wellesley.edu/Economics/mcewan/PDF/cea.pdf.

van der Tak, Herman, and Lyn Squire. 1975. *Economic Analysis of Projects*. Washington, DC: World Bank.

Warner, A. 2010. *Cost-Benefit Analysis in World Bank Projects*. Washington, DC: World Bank. https://openknowledge.worldbank.org/bitstream/handle/10986/2561/624700PUB0Cost00Box0361484B0PUBLIC0.pdf?sequence=1.

Yates, Brian T. 2015. "Cost-Benefit and Cost-Effectiveness Analyses in Evaluation Research." In *International Encyclopedia of the Social & Behavioral Sciences*, 2nd ed., edited by James D. Wright, 55–62. Amsterdam: Elsevier.

Advanced

Belfield, Clive R., A. Brooks Bowden, and Viviana Rodriguez. 2018. "Evaluating Regulatory Impact Assessments in Education Policy." *American Journal of Evaluation* 40 (3): 335–53.

Bowden, A. Brooks, Robert Shand, Clive R. Belfield, Anyi Wang, and Henry M. Levin. 2017. "Evaluating Educational Interventions That Induce Service Receipt: A Case Study Application of City Connects." *American Journal of Evaluation* 38 (3): 405–19.

Cordes, Joseph J. 2017. “Using Cost-Benefit Analysis and Social Return on Investment to Evaluate the Impact of Social Enterprise: Promises, Implementation, and Limitations.” *Evaluation and Program Planning* 64: 98–104.

Da’ar, Omar B., and Abdulaziz Alshaya. 2018. “Is It Cost-Beneficial to Society? Measuring the Economic Worth of Dental Residency Training.” *Evaluation and Program Planning* 68: 117–23.

European Commission. 2014. *Guide to Cost-Benefit Analysis of Investment Projects: Economic Appraisal Tool for Cohesion Policy 2014–2020*. Brussels: European Commission. https://ec.europa.eu/regional_policy/sources/docgener/studies/pdf/cba_guide.pdf.

Marchante, A. J., and B. Ortega. 2010. “Evaluating Efficiency in the Implementation of Structural Funds Operations.” *Evaluation XVI* (2): 193–209.

OECD. 2018. *Cost-Benefit Analysis and the Environment: Further Developments and Policy Use*. Paris: OECD Publishing.

Robinson, Lisa, ed. 2019. “Benefit-Cost Analysis in Low- and Middle-Income Countries: Methods and Case Studies.” Special issue, *Journal of Benefit-Cost Analysis* 10 (S1). https://www.cambridge.org/core/journals/journal-of-benefit-cost-analysis/issue/special-issue-benefitcost-analysis-in-low-and-middleincome-countries-methods-and-case-studies/CEA50B949FD2F37E60A8E1C0528A9112.

World Bank. 2018. *Socioeconomic Analysis of the Potential Benefits of Modernizing Hydrometeorological Services in the Lao People’s Democratic Republic*. Washington, DC: World Bank. http://documents.albankaldawli.org/curated/ar/842621563163324249/pdf/Socioeconomic-Analysis-of-the-Potential-Benefits-of-Modernizing-Hydrometeorological-Services-in-the-Lao-People-s-Democratic-Republic.pdf.

Other Resources

The Cost-Benefit & Cost-Effectiveness Hub of the Inter-American Development Bank is a one-stop shop for information and tools used for economic analysis. https://www.iadb.org/en/topics-effectiveness-improving-lives/cost-benefit-and-cost-effectiveness-resources.

BRIEF DESCRIPTION OF THE APPROACH

The primary purpose of experimental designs, commonly referred to as randomized controlled trials, is to provide an accurate estimate of (net) program effects. The defining feature of an experimental design is that people are allocated at random to either a treatment or a control group. Whereas the treatment group receives the program services, the control group receives regular or no services. The underlying logic of the random allocation is that any (observable and unobservable) differences among the treatment or the control groups are evenly distributed between the two groups. Accordingly, any observed outcome differences between the two groups can reasonably be attributed to the program being studied. In this way, experimental designs can help determine whether (and the extent to which) a cause-effect relation exists between the program and the outcome.

Evaluation questions that experimental designs may answer include the following:

1. What is the net effect of the program?
2. How does the net effect of the program vary across subgroups of the population?
3. How much do program variations affect the net effect estimate?

THE MAIN VARIATIONS OF THE APPROACH

A survey of real-world applications of experimental designs reveals a number of design variations. One variant is *multiarm designs*, where participants are randomly allocated to one of several treatment groups or one of several control groups. This design variant is useful when comparing multiple program variations (that is, multiple treatments). Another common variant is the *wait-list design* (or pipeline design), where individuals are randomly allocated to immediate program admission or to a wait-list for later program admission, allowing both for accurate effect size estimates and for all the participants to receive the treatment by the end of the evaluation.

Experimental designs may also differ according to the level of randomization. Random allocation can be at the individual level (individual people are randomly assigned to treatment or control) or cluster level (groups of people [for example, communities, districts, or schools] are randomly assigned to treatment or control).

Cluster-level randomization is often applied when the program being studied is directed at groups of people (for example, entire villages or communities), as opposed to specific individuals.

THE MAIN PROCEDURAL STEPS OF THE APPROACH

In practice, experimental designs consist of the following six steps:

- Identifying the target population for the program;
- Collecting baseline data on a representative sample of this population;
- Randomly allocating the people in the sample to either the treatment or the control group;
- Implementing the program;
- Collecting outcome data on both groups (covering at least two data points over time); and
- Comparing outcome patterns between the treatment and the control group.

Data from experimental designs can be analyzed in many different ways. Randomization allows simple mean comparisons between the treatment group and the control group (or subgroups within these) to provide an estimate of the average program effect.

Another common technique is to compare gain scores, that is, to compute the average difference between the baseline and endline outcome measures for each group and then compare these averages for a mean difference score (this is also known as the difference-in-differences approach, which is described in more detail in guidance note 3, Quasi-Experimental Approaches).

If the treatment and control group (despite randomization) differ on baseline characteristics, statistical techniques such as multiple regression analysis can be applied to adjust for these differences when estimating the program effect. Finally, when combined with information on program costs, data from experimental designs can also support retrospective cost-effectiveness and cost-benefit analysis (see guidance note 1, Efficiency Analysis: Cost-Benefit and Cost-Effectiveness).

THE ADVANTAGES AND DISADVANTAGES OF THE APPROACH

An experimental design is particularly relevant when answering evaluation questions related to program effectiveness, particularly the net effect of the program on a specific outcome or set of outcomes. If implemented well, the experimental design provides the most accurate estimate of the net program effect on selected outcomes. The reliance on random assignment enhances the internal validity of any causal claims produced by experimental designs (see *internal validity* in appendix A, Glossary of Key Terms). In this way, the design serves well to establish cause-effect relationships between a program and its outcomes.

Experimental designs also come with several practical and ethical challenges. First, the design relies on a very stable program implementation and a homogeneous target group to provide accurate program effect estimates. However, these conditions are difficult to maintain in practice and may even reduce the generalizability (external validity) of the evaluation findings (see *external validity* in appendix A, Glossary of Key Terms). Second, estimates significantly improve with multiple data points. Pathways of change may be nonlinear and two data points (for example, before and after only) may be too limited for reliably estimating the net effect. A third common methodological concern is the possibility of contamination. Contamination can arise from the program itself as a result of spillover effects from individuals in the treatment and the control group influencing each other (own contamination). A fourth concern is that the focus of the analysis is restricted to one or more measurable intended effects. Consequently, it is less suitable for assessing unintended effects. Finally, experimental designs are accurate only if two conditions are met: (i) The evolution or development of treatment and control groups is homogeneous throughout the intervention implementation. This includes homogeneity of intervention across the treatment group. Otherwise, emerging systematic differences between the two groups may result in bias when the program effects are estimated. (ii) A certain level of statistical power is needed to reach statistical significance, under which the findings are not reliable.

In addition to these practical challenges, the ethical implications of withholding program activities from people in the control group may pose a barrier; the use of wait-list (pipeline) designs, however, may alleviate this concern.

THE APPLICABILITY OF THE APPROACH

When considering the use of experimental designs, evaluators must plan for the randomized assignment of treatment and control groups, and the collection of baseline data for comparison with data collected later. The evaluator therefore should be involved in the design and implementation stages of the evaluation. The evaluators also need to plan and budget for the time- and resource-consuming task of collecting data on both the treatment and the control group. Many types of evaluators will be involved in these early stages of an evaluation, including impact evaluation experts (typically researchers), internal agency evaluators, and external evaluators hired for this purpose. Yet, because of this early and direct engagement with the intervention, a typical IEO evaluator is very unlikely to be involved directly in an experimental design. Nevertheless, though impact evaluations are not typically managed by IEOs, credible randomized controlled trial studies can be very helpful as inputs to other types of retrospective studies on those same programs or as part of a sectorwide evaluation, as sources of evidence for a structured or systematic review. In fact, in lieu of direct involvement in impact evaluations, IEOs, such as the World Bank's Independent Evaluation Group, do perform systematic reviews (or variations thereof), which draw on impact evaluations carried out by others to deepen the evidence base for sector and thematic evaluations (see guidance note 11, Structured Literature Reviews).

Despite their limitations, applications of experimental designs have over the years covered a broad range of programs and sectors in development evaluation, including the following:

1. An impact evaluation of an adolescent development program for girls in Tanzania used a multiarm experimental design.

   (*Source*: Buehran, N., M. Goldstein, S. Gulesci, M. Sulaiman, and V. Yam. 2017. "Evaluation of an Adolescent Development Program for Girls in Tanzania." Policy Research Working Paper 7961, World Bank, Washington, DC. http://documents.worldbank.org/curated/en/245071486474542369/pdf/WPS7961.pdf.)

2. An experiment was used in the impact evaluation of a mobile point of service deposit collection for business owners in Sri Lanka. Self-employed individuals were randomly allocated to a treatment program offering weekly door-to-door savings deposit collection services and assistance opening bank accounts.

(*Source*: Callen, M., C. McIntosh, S. de Mel, and C. Woodruff. 2014. “What Are the Headwaters of Formal Savings? Experimental Evidence from Sri Lanka.” NBER Working Paper 20736, National Bureau of Economic Research, Cambridge, MA. https://www.povertyactionlab.org/evaluation/impact-formal-savings-intervention-sri-lanka.)

3. An experiment was used to better understand the influence of psychic and economic barriers on vaccination rates.

   (*Source*: Sato, R., and Y. Takasaki. 2018. “Psychic vs. Economic Barriers to Vaccine Take-Up: Evidence from a Field Experiment in Nigeria.” Policy Research Working Paper 8347, World Bank, Washington, DC. http://documents.worldbank.org/curated/en/876061519138798752/pdf/WPS8347.pdf.)

4. Randomized experiments were used in the evaluation of the Balsakhi remedial education program in Mumbai, India.

   (*Source*: Banerjee, A., S. Cole, E. Duflo, and L. Lindon. 2007. “Remedying Education: Evidence from Two Randomized Experiments in India.” *Quarterly Journal of Economics* 122 (3): 1235–64.)

5. The pipeline design using cluster (community) randomized allocation was applied in an evaluation of a conditional cash transfer program (the PROGRESA/Oportunidades program).

   (*Source*: Bamberger, M., and A. Kirk. 2009. *Making Smart Policy: Using Impact Evaluation for Policy-Making: Case Studies on Evaluations That Influenced Policy*. Doing Impact Evaluation 14. Washington, DC: World Bank. http://documents.worldbank.org/curated/en/239681468324546563/Making-smart-policy-using-impact-evaluation-for-policy-making-case-studies-on-evaluations-that-influenced-policy.)

6. Random allocation of interest rates was used in an evaluation of a loan and consumer credit program in South Africa.

   (*Source*: Karlan, D., and J. Zinman. 2003. “Interest Rates and Consumer Credit in South Africa.” Study Summary. New Haven, CT: Innovations for Poverty Action. https://www.poverty-action.org/printpdf/6326.)

7. An experiment was used in the evaluation of a rural microfinance program on agricultural and nonagricultural activities, income, and expenditures in Morocco.

   (*Source*: Bamberger, M., and A. Kirk. 2009. *Making Smart Policy: Using Impact Evaluation for Policy-Making: Case Studies on Evaluations That Influenced Policy*. Doing Impact Evaluation 14. Washington, DC: World Bank. http://documents.worldbank.org/curated/en/239681468324546563/Making-smart-policy-using-impact-evaluation-for-policy-making-case-studies-on-evaluations-that-influenced-policy.)

8. An experiment was used to evaluate the effectiveness of insecticide-treated bed nets for malaria prevention in Kenya.

   (*Source*: Bamberger, M., and A. Kirk. 2009. *Making Smart Policy: Using Impact Evaluation for Policy-Making: Case Studies on Evaluations That Influenced Policy*. Doing Impact Evaluation 14. Washington, DC: World Bank. http://documents.worldbank.org/curated/en/239681468324546563/Making-smart-policy-using-impact-evaluation-for-policy-making-case-studies-on-evaluations-that-influenced-policy.)

9. An experiment was used in the evaluation of three mother-literacy interventions in rural India to improve child learning through increased mother literacy and direct encouragement of learning at home. Villages were randomly allocated to mother literacy interventions.

   (*Source*: Banerji, R., J. Berry, and M. Shortland. 2014. *The Impact of Mother Literacy and Participation Programmes on Child Learning: Evidence from a Randomised Evaluation in India*. 3ie Impact Evaluation Report 16. New Delhi: International Initiative for Impact Evaluation. https://www.3ieimpact.org/evidence-hub/publications/impact-evaluations/impact-mother-literacy-and-participation-programmes.)

10. An experiment was used in an impact evaluation of a voucher program for out-of-school youth, to measure earnings, including wage earnings; self-employed profits; and labor market outcomes. Individual youth were randomly allocated to treatment in the form of a voucher for vocational training.

   (*Source*: Hamory, J., M. Kremer, I. Mbiti, and E. Miguel. 2016. *Evaluating the Impact of Vocational Education Vouchers on Out-Of-School Youth in Kenya*. 3ie Impact Evaluation Report 37. New Delhi: International

Initiative for Impact Evaluation. https://www.3ieimpact.org/evidence-hub/publications/impact-evaluations/evaluating-impact-vocational-education-vouchers-out.)

READINGS AND RESOURCES

Background

Duflo, Esther, Rachel Glennerster, and Michael Kremer 2007. "Using Randomization in Development Economics Research: A Toolkit." Center for Economic Policy Research Discussion Paper 6059, Massachusetts Institute of Technology, Cambridge, MA. https://economics.mit.edu/files/806.

Gertler, P. J., S. Martinez, P. Premand, L. B. Rawlings, and C. M. J. Vermeersch. 2016. *Impact Evaluation in Practice*, 2nd ed. Washington, DC: Inter-American Development Bank and World Bank. https://openknowledge.worldbank.org/handle/10986/25030.

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

Khandker, S. R., G. B. Koolwal, and H. A. Samad. 2009. *Handbook on Quantitative Methods of Program Evaluation*. Washington, DC: World Bank. http://documents.worldbank.org/curated/en/650951468335456749/pdf/520990PUB0EPI1101Official0Use0Only1.pdf.

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

Advanced

Bell, S. H., and L. R. Peck. 2016. "On the 'How' of Social Experiments: Experimental Designs for Getting Inside the Black Box." *New Directions for Evaluation* 152: 97–107.

Faulkner, W. N. 2014. "A Critical Analysis of a Randomized Controlled Trial Evaluation in Mexico: Norm, Mistake or Exemplar?" *Evaluation* 20 (2): 230–43.

Glewwe, Paul, and Petra Todd, eds. 2020. *Impact Evaluation in Developing Countries: Theory, Methods, and Practice*. Washington, DC: World Bank.

Higuchi, Yuki, Edwin P. Mhede, and Tetsushi Sonobe. 2019. "Short- and Medium-Run Impacts of Management Training: An Experiment in Tanzania." *World Development* 114: 220–36.

Ledford, Jennifer R. 2018. "No Randomization? No Problem: Experimental Control and Random Assignment in Single Case Research." *American Journal of Evaluation* 39 (1): 71–90.

McCarthy, Aine Seitz. 2019. “Intimate Partner Violence and Family Planning Decisions: Experimental Evidence from Rural Tanzania.” *World Development* 114: 156–74.

Tipton, E., and B. J. Matlen. 2019. “Improved Generalizability through Improved Recruitment: Lessons Learned from a Large-Scale Randomized Trial.” *American Journal of Evaluation* 40 (3): 414–30.

Other Resources

The Abdul Latif Jameel Poverty Action Lab is a global research center working to reduce poverty by ensuring that policy is informed by scientific evidence. Anchored by a network of 194 affiliated professors at universities around the world, it conducts randomized impact evaluations to answer critical questions in the fight against poverty. https://www.povertyactionlab.org/about-j-pal; https://www.poverty-action.org/about/randomized-control-trials.

The International Initiative for Impact Evaluation funds, produces, quality assures, and synthesizes evidence of policy interventions that work in low- and middle-income countries and advocates the generation and use of quality evidence in development decision-making. https://www.3ieimpact.org/; https://developmentevidence.3ieimpact.org/.

The World Bank’s Development Impact Evaluation group generates high-quality and operationally relevant data and research to transform development policy, help reduce extreme poverty, and secure shared prosperity. https://www.worldbank.org/en/research/dime; https://dimewiki.worldbank.org/wiki/Main_Page.

3 Quasi-Experimental Approach

BRIEF DESCRIPTION OF THE APPROACH

Like experimental designs, quasi-experimental designs are meant to provide an accurate estimate of (net) program effects. The main difference involves random assignment. Although random assignment is fundamental for experimental design, quasi-experiments do *not* rely on random assignment of people to establish treatment or comparison groups. Instead, quasi-experiments rely on a broad range of statistical techniques to construct treatment and comparison groups that are comparable in terms of a select set of baseline characteristics. For quasi-experimental designs, the term *comparison group* is often used instead of *control group*, which is the term used in experimental design with randomization.

As for randomized designs, evaluation questions that quasi-experimental designs may answer include the following:

1. What is the net effect of the program?
2. How does the net effect of the program vary across subgroups of the population?
3. How much do programmatic variations affect the net effect estimate?

THE MAIN VARIATIONS OF THE APPROACH

There are several quasi-experimental designs, and four common types are described here.

In *propensity score matching*, people in the treatment group are matched with comparable people (sometimes referred to as “twins”) in the comparison group. The matching is based on the (observable) characteristics of the population believed to affect the probability of participating in the program, summarized in an overall score representing their propensity to be enrolled. The common support (or overlapping) interval represents the range of propensity scores for which both enrolled and unenrolled units are available. The outcomes observed in these two groups are then compared to estimate the program effect. Matching may be applied at the individual or group level; for example, students could be matched with other students, or schools could be matched with other schools.

In *regression discontinuity designs*, a program eligibility criterion (for example, income level or test score) is used to construct comparable groups (ideally the program eligibility criterion is not associated with other benefits—for example, other state benefits). The core idea of regression discontinuity design is that individuals immediately below the program cut-off score (those who were not accepted into the program) are similar to those immediately above the cut-off (those who were accepted). To illustrate, consider a program where rural farmers above a specific income level are eligible for a tractor lease program. Those farmers just below the cut-off (the comparison group), although not admitted to the program, are likely to be comparable to those farmers immediately above the cut-off (the treatment group). A regression-based comparison of the difference in average outcomes for these two groups can be used to estimate the program effect.

The *instrumental variable* method uses a variable that is correlated with program participation (but not with the program outcome of interest) to adjust for factors affecting the likelihood of program participation. The program effect is then estimated using a regression model containing the instrumental variable, among other relevant covariates.

The *difference-in-differences* method estimates the program effect by comparing the difference over time among nonparticipants with that among program participants (that is, the difference in the differences). This approach eliminates external determinants of the outcome that are time-invariant for the treatment and comparison group during the program period.

THE MAIN PROCEDURAL STEPS OF THE APPROACH

This guide provides only broad descriptions of these four quasi-experimental approaches. Additional reading on these topics is advised for readers considering them. The basic steps for propensity score matching and regression discontinuity are presented in this section. But the steps for difference-in-differences and instrumental variable approaches require explanation of statistical and analytical steps that are outside of the scope of this guide. Many resources are readily available to describe the methods and steps, such as those from the International Initiative for Impact Evaluation, listed in Other Resources.

The procedural steps of quasi-experimental designs vary according to the way in which the treatment and comparison groups are constructed.

Propensity score matching generally involves the following five steps:

- Making assumptions on the factors affecting participation in the program;
- Modeling the relevant variables with a logistic regression model explaining participation or exclusion;
- Estimating the propensity to participate in the program (for participants and nonparticipants);
- Matching participants and nonparticipants sharing similar propensity scores; and
- Comparing their evolution over the course of the program and thus estimating program effects.

The regression discontinuity design consists of four steps:

- Identifying the margin around the cut-off score for program participation where individuals are comparable;
- Fitting a regression line on these individuals' cut-off scores and outcome scores;
- Identifying any shift (discontinuity) in the regression line at the cut-off score; and
- Interpreting the size of the shift as the estimated program effect.

THE ADVANTAGES AND DISADVANTAGES OF THE APPROACH

Quasi-experimental designs are particularly relevant when the evaluation emphasis is on program effectiveness, and random assignment of people to treatment and control is not possible. In these situations, the quasi-experimental designs may provide the least biased program effect estimates, as compared with, for instance, nonexperimental designs that usually have no or less robust comparison arrangements (see the Maryland Scientific Methods Scale in Other Resources for a ranking of treatment-comparison designs). Moreover, some quasi-experimental designs (for example, propensity score matching) can also be used retrospectively, that is, after the program has been implemented. However, baseline data are usually preferred. Many quasi-experimental designs, however, are attractive to evaluators who may have access to the data from an intervention but have no opportunity for direct involvement with the intervention (that is, in collecting baseline or endline data directly).

Quasi-experiments are not without shortcomings. One methodological weakness of quasi-experimental designs emerges from the lack of random assignment, potentially resulting in treatment and comparison groups that are different in ways that may affect the estimated program effects. Because the construction of comparable groups solely by statistical means accounts for observable characteristics (or time-invariant unobservable differences), the extent to which estimates of program effects are influenced by unobserved differences is a persistent concern (see *selection bias* in appendix A, Glossary of Key Terms). Again, much depends on the availability of data and the number of data points for both treatment and comparison groups over time. Finally, even when the design is solid and the comparison group result is unbiased, accurate, and comparable with the treatment group result, the data might not be sufficiently precise, because a certain level of statistical power is needed to reach statistical significance.

THE APPLICABILITY OF THE APPROACH

The lack of random assignment is typical of most development programs. Quasi-experiments are in practice more applicable in development contexts than experimental designs. However, other factors, including the requirement for baseline data for most quasi-experimental designs, make the approach less applicable to IEO operations. A further limit to the practical application of quasi-experimental designs is the time- and resource-consuming task of collecting data on both the program and the comparison groups.

Applications of quasi-experimental design include the following:

1. Propensity score matching was used to identify villages that were similar in socioeconomic terms as part of an evaluation of a conflict resolution program (Kecamatan) in Indonesia.

   (*Source*: Voss, John. 2008. "Impact Evaluation of the Second Phase of the Kecamatan Development Program in Indonesia." Working Paper 45590, World Bank, Washington, DC. http://documents.worldbank.org/curated/en/551121468048909312/Impact-evaluation-of-the-second-phase-of-the-Kecamatan-development-program-in-Indonesia.)

2. Propensity score matching was used to match households (on background variables) as part of an impact evaluation on water and sanitary interventions in Nepal.

   (*Source*: Bose, R. 2009. "The Impact of Water Supply and Sanitation Interventions on Child Health: Evidence from DHS Surveys." Paper presented at the Biannual Conference on Impact Evaluation, Colombo, Sri Lanka, April 22–23. https://editorialexpress.com/cgi-bin/conference/download.cgi?db_name=FEMES09&paper_id=204.)

3. Propensity score matching combined with a difference-in-differences method was used to estimate the effect of a conditional cash transfer program in Chile that sought to improve several socioeconomic outcomes for families living in poverty.

   (*Source*: Martorano, Bruno, and Marco Sanfilippo. 2012. "Innovative Features in Conditional Cash Transfers: An Impact Evaluation of Chile Solidario on Households and Children." Innocenti Working Paper 2012–03, UNICEF Innocenti Research Centre, Florence. https://www.unicef-irc.org/publications/656-innovative-features-in-conditional-cash-transfers-an-impact-evaluation-of-chile-solidario.html.)

4. A regression discontinuity design was used in the evaluation of an educational program under the PROGRESA poverty alleviation program in Mexico City, where children were admitted on the basis of a household income index score.

   (*Source*: Buddelmeyer, Hielke, and Emmanuel Skoufias. 2004. "An Evaluation of the Performance of Regression Discontinuity Design on PROGRESA." Policy Research Working Paper WPS 3386, World Bank, Washington, DC. http://documents.worldbank.org/curated/en/124781468773348613/An-evaluation-of-the-performance-of-regression-discontinuity-design-on-PROGRESA.)

5. An instrumental variable approach was used in a World Bank evaluation of an energy-efficiency program in Ethiopia that distributed compact fluorescent lamp bulbs free of charge to poor households.

   (*Source*: Costolanski, P., R. Elahi, A. Limi, and R. Kitchlu. 2013. "Impact Evaluation of Free-of-Charge CFL Bulb Distribution in Ethiopia." Policy Research Working Paper 6383, World Bank, Washington, DC. http://documents1.worldbank.org/curated/en/294421468032720712/pdf/wps6383.pdf.)

6. An evaluation of the impact of minimum wages on employment used matched difference-in-differences estimates of the employment impact in select industries in Indonesia.

(*Source*: Alatas, V., and L. A. Cameron. 2003. "The Impact of Minimum Wages on Employment in a Low-Income Country: An Evaluation Using Difference-in-Differences Approach." Policy Research Working Paper 2985, World Bank, Washington, DC.)

7. The instrumental variable approach was applied to evaluate the impact of infrastructure development on economic growth and income distribution in Latin American countries.

   (*Source*: Calderon, C., and L. Serven. 2004. "The Effects of Infrastructure Development on Growth and Income Distribution." Policy Research Working Paper 3400, World Bank, Washington, DC. http://documents1.worldbank.org/curated/en/438751468753289185/pdf/WPS3400.pdf.)

8. An impact evaluation of a World Bank Credit Program on small and medium enterprises in Sri Lanka used propensity score matching for measuring program impact.

   (*Source*: Aivazian, V. A., and E. Santor. 2008. "Financial Constraints and Investment: Assessing the Impact of a World Bank Credit Program on Small and Medium Enterprises in Sri Lanka." *Canadian Journal of Economics* 41 (2): 475–500.)

9. A regression discontinuity design was used in the evaluation of Burkinabé Response to Improve Girls' Chances to Succeed, a two-year program aimed at improving girls' access to primary school.

   (*Source*: Levy, D., M. Sloan, L. Linden, and H. Kazianga. 2009. *Impact Evaluation of Burkina Faso's BRIGHT Program*. Washington, DC: Mathematica Policy Research. https://eric.ed.gov/?id=ED507466.)

10. A quasi-experimental design with propensity score–matched comparison villages was used in an impact assessment of a value chain development of bay leaf in Nepal.

    (*Source*: Shah, G. M., A. K. Nepal, G. Rasul, and F. Ahmad. 2018. "Value Chain Development of Bay Leaf in Nepal: An Impact Assessment." *Journal of Development Effectiveness* 10 (2): 179–96.)

11. A propensity score matching method was used to adjust for baseline differences in a randomized controlled trial, comparing microfinance institution borrowers to those without any loans and those with loans from other sources.

(*Source*: Inna, C., and I. Love. 2017. “Re-evaluating Microfinance—Evidence from a Propensity-Score Matching.” Policy Research Working Paper 8028, World Bank, Washington, DC. http://documents.worldbank.org/curated/en/707891492451043846/Re-evaluating-microfinance-evidence-from-propensity-score-matching.)

READINGS AND RESOURCES

Background

Gertler, P. J., S. Martinez, P. Premand, L. B. Rawlings, and C. M. J. Vermeersch. 2016. *Impact Evaluation in Practice*, 2nd ed. Washington, DC: World Bank. https://openknowledge.worldbank.org/handle/10986/25030.

Khandker, S. R., G. B. Koolwal, and H. A. Samad. 2009. *Handbook on Quantitative Methods of Program Evaluation*. Washington, DC: World Bank.

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

White, H., and S. Sabarwal. 2014. “Quasi-Experimental Design and Methods.” Methodological Briefs: Impact Evaluation 8, UNICEF Office of Research, Florence. https://www.unicef-irc.org/publications/753-quasi-experimental-design-and-methods-methodological-briefs-impact-evaluation-no.html.

Advanced

Buddelmeyer, Hielke, and Emmanuel Skoufias. 2004. “An Evaluation of the Performance of Regression Discontinuity Design on PROGRESA.” Policy Research Working Paper WPS 3386, World Bank, Washington, DC. http://documents.worldbank.org/curated/en/124781468773348613/An-evaluation-of-the-performance-of-regression-discontinuity-design-on-PROGRESA.

Caliendo, M., and S. Kopeinig. 2008. “Some Practical Guidance for the Implementation of Propensity Score Matching.” *Journal of Economic Surveys* 22: 31–72. http://ftp.iza.org/dp1588.pdf.

Gao, Xingyuan, Jianping Shen, and Huilan Y. Krenn. 2017. “Using WWC Sanctioned Rigorous Methods to Develop Comparison Groups for Evaluation.” *Evaluation and Program Planning* 65: 148–55.

Glewwe, Paul, and Petra Todd, eds. 2020. *Impact Evaluation in Developing Countries: Theory, Methods, and Practice*. Washington, DC: World Bank.

Hansen, H., N. Klejnstrup, and O. W. Andersen. 2013. “A Comparison of Model-Based and Design-Based Impact Evaluations in Developing Countries.” *American Journal of Evaluation* 34 (3): 320–38.

Mourelo, Elva López, and Verónica Escudero. 2017. "Effectiveness of Active Labor Market Tools in Conditional Cash Transfers Programs: Evidence for Argentina." *World Development* 94: 422–47.

Weitzman, B. C., D. Silver, and K-N. Dillman. 2002. "Integrating a Comparison Group Design into a Theory of Change Evaluation: The Case of the Urban Health Initiative." *American Journal of Evaluation* 23 (4): 371–86.

Wing, Coady, and Ricardo A. Bello-Gomez. 2018. "Regression Discontinuity and Beyond: Options for Studying External Validity in an Internally Valid Design." *American Journal of Evaluation* 39 (1): 91–108.

Other Resources

The International Initiative for Impact Evaluation funds, produces, assures quality, synthesizes evidence of what policy interventions work in low- and middle-income countries, and advocates the generation and use of quality evidence in development decision-making. https://www.3ieimpact.org/; https://developmentevidence.3ieimpact.org/.

The Maryland Scientific Methods Scale is a five-point scale ranking treatment-comparison designs according to their power to reduce selection bias. https://whatworksgrowth.org/resources/the-scientific-maryland-scale/.

The World Bank's Development Impact Evaluation group generates high-quality and operationally relevant data and research to transform development policy, help reduce extreme poverty, and secure shared prosperity. https://www.worldbank.org/en/research/dime; https://dimewiki.worldbank.org/wiki/Main_Page.

4 Case Study Design

BRIEF DESCRIPTION OF THE APPROACH

The case study approach is a focused, in-depth examination of one or more specific and clearly defined cases (individuals, programs, organizations, communities, or even countries). The purpose of case studies in evaluation is often to explore and better understand how a program was implemented and to identify causal processes and configurations generating program outcomes, including contextual factors conditioning these. Case studies are particularly adept at documenting program contextual conditions, substantiating how and in what way the program generated (or failed to generate) one or more intended outcomes, and even producing insights about whether and how the program might make a difference in other settings, times, and populations (see *analytical generalization* in appendix A, Glossary of Key Terms). *Case study* is an umbrella term comprising several different design subtypes, many of which are discussed in this guide: process tracing, qualitative comparative analysis, participatory approaches, and (some) complex systems approaches are all case-based approaches and usually applied to a handful of cases at most.

Many types of evaluation questions can be answered by case studies, including the following:

1. How was the program implemented?
2. How and in what way did the program generate the observed effect?
3. Will the program make a difference in other settings, times, and populations?

A case study design in principle can be used for any type of evaluation question and is not restricted to causal questions.

THE MAIN VARIATIONS OF THE APPROACH

Case studies can be designed and implemented in many different ways. Some case study designs center on a single case; others include multiple cases. Case studies can be implemented at a single point in time or repeated over time (for example, before, during, and after program implementation). In terms of data collection, case studies may involve and be greatly improved by a broad range of qualitative and quantitative methods, including combinations of these.

Case studies may be further distinguished in purpose as illustrative (to describe a typical or abnormal case), theory-generating (to explore and generate hypotheses), theory testing (to test and revise hypotheses), and cumulative (to compare and synthesize multiple cases).

Case study analyses may include within-case analysis, cross-case analysis, or some combination of these (see guidance notes 5, Process Tracing, and 6, Qualitative Comparative Analysis, for examples). There are many ways that case study data can be examined. These include strategies for identifying similarities and differences across cases; multiple data display tables for partitioning and grouping data in various ways, sometimes based on features of the included cases or time-ordered displays; and coding techniques for further qualitative or even statistical analyses.

THE MAIN PROCEDURAL STEPS OF THE APPROACH

A case study typically involves the following five steps:

- Identifying and defining the type of case(s) to be examined (this is also referred to as casing);
- Identifying the conditions or factors that will be studied in more depth in these case(s);
- Developing a case selection strategy;
- Collecting data on the selected case(s); and
- Analyzing the data using within-case or cross-case analytical techniques, or some combination of these.

A central step in case study design, and one that in practice is all too often treated less carefully than it deserves, is the selection of the case or the cases to be included (step iii above). The purpose of the case study (and the type of information to be produced by it) should inform the selection of relevant cases. To illustrate, if the aim of the case study is to gauge the breadth of and variation among local program implementation processes, the case selection should focus on the program implementation processes considered most different on a set of relevant characteristics (for example, urban versus rural, small- versus large-scale cases). Conversely, if the aim of the case study is to better understand high-performing programs, a better case selection strategy could be to focus on these programs (as defined by one or more program goals). Finally, random selection is often inappropriate for case study selection, partly because the number of cases tends to be too low for randomization to balance out systematic differences.

THE ADVANTAGES AND DISADVANTAGES OF THE APPROACH

There are several important advantages to case studies. Emphasizing in-depth analysis, case studies can identify and examine causal processes underlying programs (also known as mechanisms) and the context within which these processes are embedded (see guidance note 5, Process Tracing). In this way, case studies may generate hypotheses about these underlying processes, examine these in more detail within a single case or across multiple cases, and even identify the specific contextual conditions on which these processes are contingent.

Another notable strength of case studies, particularly multiple–case study designs, is the opportunity to generalize findings beyond the programs studied (that is, these cases) to other programs that are similar on one or more salient characteristics (see *analytical generalization* in appendix A, Glossary of Key Terms).

There are also limitations. Commonly cited limitations include the absence of clear procedural guidelines for different variants of cases studies, the vulnerability of case studies to evaluator subjectivity (in part from lack of procedural formalization of case studies), the limited generalizability (of single–case study designs—see chapter 2 for a discussion of the trade-offs between breadth and depth), and the practical challenge of time and effort needed to adequately carry out case studies (especially for mixed data collection methods across multiple cases). Finally, the case study design is also vulnerable to the analyst "cherry-picking" specific cases to find support for preestablished ideas about the program or to present the most dramatic cases (the poorest family, the most successful entrepreneur) rather than to present a broader range of cases.

THE APPLICABILITY OF THE APPROACH

The case study design can be applied in most settings and contexts. As such, case studies are highly applicable in development evaluation in general and IEO evaluations specifically. Case studies may also complement other approaches and designs that tend to focus less on contextual conditions and how these interact with the program (for example, experimental or quasi-experimental designs).

Case studies are widely used in development evaluation; examples include the following:

1. Case studies and other approaches were used in an evaluation of citizen engagement mainstreaming in World Bank projects.

(*Source*: World Bank. 2018. *Engaging Citizens for Better Development Results*. Independent Evaluation Group. Washington, DC: World Bank. https://ieg.worldbankgroup.org/evaluations/engaging-citizens-better-development-results.)

2. Multiple case studies, anchored in and framed by a program theory, were used to enhance the generalizability of findings from an evaluation of the Africa Routine Immunization program. This is a good example of theory-informed case selection and use of within- and across-case analyses.

   (*Source*: Mookherji, S., and A. LaFond. 2013. "Strategies to Maximize Generalization from Multiple Case Studies: Lessons from the Africa Routine Immunization System Essentials (ARISE) Project." *Evaluation* 19 (3): 284–303.)

3. Six individual cases studies of cities (Bucaramanga, Colombia; Coimbatore, India; Kigali, Rwanda; Gaziantep, Turkey; Changsha, China; and Tangier, Morocco) were compared with each other to identify institutions and strategies that successful cities have relied on to spur economic development.

   (*Source*: Kulenovic, Z. J., and A. Cech. 2015. "Six Case Studies of Economically Successful Cities: Competitive Cities for Jobs and Growth." Companion Paper 3, Washington, DC, World Bank. https://openknowledge.worldbank.org/handle/10986/23573.)

4. A case study approach was used in a poverty and social impact evaluation of Tanzania's crop boards reform.

   (*Source*: Beddies, S., M. Correia, S. Kolavalli, and R. Townsend. 2006. "Tanzania Crop Boards Reform." In *Poverty and Social Impact Analysis of Reforms: Lessons and Examples from Implementation*, edited by A. Coudouel, A. Dani, and S. Paternostro, 491–520. Washington, DC: World Bank. http://regulationbodyofknowledge.org/wp-content/uploads/2014/09/Coudouel_Poverty_and_Social.pdf.)

5. A case study approach was used to assess the impact of a government decision to close down subsidized wheat flour ration shops intended to provide wheat flour to low-income groups in Pakistan.

(*Source*: World Bank 2005. *Influential Evaluations: Detailed Case Studies*. Washington, DC: World Bank. http://documents.worldbank.org/curated/en/928001468330038416/pdf/328800Influent1luation1case1studies.pdf.)

6. Case studies were used in the evaluation of the Children of Uruzgan Program in Afghanistan. The in-depth examination of the program development generated insights to inform program planning, decision-making, and scale-up.

   (*Source*: Save the Children Australia. 2012. *Access Restricted: A Review of Remote Monitoring Practices in Uruzgan Province*. East Melbourne, Victoria: Save the Children Australia. https://resourcecentre.savethechildren.net/node/8291/pdf/access-restricted-save-the-children.pdf.)

7. A case study design was used to conduct a preliminary evaluation of the potential costs and benefits of rehabilitation of the Nakivubo wetland, Kampala, Uganda.

   (*Source*: Turpie, J., L. Day, D. Gelo Kutela, G. Letley, C. Roed, and K. Forsythe. 2016. *Promoting Green Urban Development in Africa: Enhancing the Relationship between Urbanization, Environmental Assets and Ecosystem Services*. Washington, DC: World Bank. https://openknowledge.worldbank.org/handle/10986/26425.)

8. A multiple–case study design of 14 water and sanitation policy initiatives (across seven countries) was used to identify when, why, and how sanitation utilities can work together toward specific policy outcomes.

   (*Source*: Kis, A. L., and M. Salvetti. 2017. *Case Study—Alföldvíz, Hungary*. WSS GSG Utility Turnaround Series, World Bank, Washington, DC. https://openknowledge.worldbank.org/handle/10986/27982.)

READINGS AND RESOURCES

Background

Box-Steffensmeier, Janet M., Henry E. Brady, and David Collier, eds. 2008. "Qualitative Tools for Descriptive and Causal Inference." In *The Oxford Handbook of Political Methodology*. Oxford: Oxford University Press.

Collier, David. 1993. “The Comparative Method.” In *Political Science: The State of Discipline II*, edited by Ada W. Finifter, 105–19. Washington, DC: American Political Science Association. https://ssrn.com/abstract=1540884.

George, Alexander L., and Andrew Bennett 2005. *Case Studies and Theory Development in the Social Sciences*. Cambridge, MA: MIT Press.

Gerring, John 2017. *Case Study Research: Principles and Practices*. Cambridge, UK: Cambridge University Press.

Ragin, Charles C., and Howard Saul Becker. 1992. *What Is a Case? Exploring the Foundations of Social Inquiry*. Cambridge, UK: Cambridge University Press

Stake, R. E. 2006. *Multiple Case Study Analysis*. New York: Guilford Press.

Yin, R. K. 2017. *Case Study Research and Applications: Design and Methods*, 6th ed. Thousand Oaks, CA: SAGE.

Advanced

Brady, Henry E., and David Collier. 2010. *Rethinking Social Inquiry: Diverse Tools, Shared Standards*. Lanham, MD: Rowman & Littlefield.

Gerring, John. 2004. “What Is a Case Study and What Is It Good For?” *The American Political Science Review* 98 (2): 341–54.

Hadfield, Mark, and Michael Jopling. 2018. “Case Study as a Means of Evaluating the Impact of Early Years Leaders: Steps, Paths and Routes.” *Evaluation and Program Planning* 67: 167–76.

King, Gary, Robert O. Keohane, and Sidney Verba. 1994. *Designing Social Inquiry: Scientific Inference in Qualitative Research*. Princeton University Press.

USAID (US Agency for International Development). 2013. “Evaluative Case Studies.” Technical Note. November, USAID, Washington, DC. https://usaidlearninglab.org/sites/default/files/resource/files/case_study_tech_note_final_2013_1115.pdf.

Woolcock, M. 2013. “Using Case Studies to Explore the External Validity of ‘Complex’ Development Interventions.” *Evaluation* 19 (3): 229–48.

BRIEF DESCRIPTION OF THE APPROACH

Process tracing is a case-based approach for examining and describing the causal processes (also referred to as mechanisms) generating program outcomes. Its purpose is to identify and empirically test the causal mechanisms connecting specific program components and a set of desired outcomes within a single case. Examining these processes supports fine-grained explanations of how and why programs generate specific outcomes. What differentiates this approach from most other theory-based ones is its focus on the assessment of evidence strength—the importance or value of given data and empirical observations to support or weaken specific theories. In this sense, process tracing can be considered a data analysis technique, particularly in its explicitly Bayesian version (see The Main Variations of the Approach section). Process tracing is conventionally applied in within-case analysis of single–case study designs.

Evaluation questions that process tracing may answer include the following:

1. How, under what circumstances, and why did the program generate the desired outcome(s)?
2. Which mechanism(s) generated the desired outcome(s)?
3. Are there alternative explanations for the desired outcome(s)?

THE MAIN VARIATIONS OF THE APPROACH

Two main variations of process tracing have been defined: a traditional, purely qualitative one and a qualitative-quantitative variant where confidence levels are formally defined and updated with the Bayes formula (known sometimes as contribution tracing, Bayesian process tracing, process tracing with Bayesian updating, and, more recently, diagnostic evaluation). A secondary distinction concerns whether the investigation is conducted primarily from an inductive (theory-building) or deductive (theory-testing) perspective.

Theory-building process tracing aims to identify and describe causal processes through an empirical case. The core idea of the approach is simply to provide the best possible explanation of how the program outcomes came about within a specific local context.

Theory-testing process tracing aims to test whether specified causal processes are supported by an empirical case. This approach may serve well to identify whether a specific (theoretically derived) mechanism is present in a case being studied within a specific local context.

Bayesian formalization can coexist with any process tracing variant and also any theory-based evaluation approach, hence the more general term *diagnostic (theory-based) evaluation*. Although conceptually distinct, in practice, theory testing and theory building are usually combined, and there is often a back-and-forth exchange between theory and data across multiple iterations. Conversely, the distinction between traditional process tracing and the Bayesian version has considerable practical implications because the latter requires explicit establishment of confidence levels for a series of events concerning the theory and the empirical data in relation to the theory. Recently, process tracing has also been applied in multiple–case study designs.

THE MAIN PROCEDURAL STEPS OF THE APPROACH

There are four major steps in process tracing or diagnostic evaluation:

- Formulating the hypothesized mechanisms for the outcome;
- Specifying the observable data (in the form of patterns, timelines, traces, or accounts) that must be collected and analyzed to test for the presence of the mechanisms;
- Collecting relevant data; and
- Analyzing the collected data, considering the relative weight of evidence for the presence or absence of each hypothesized mechanism. In the fourth step, the evidence for each hypothesized mechanism is assessed using four tests (the straw-in-the-wind test, the hoop test, the smoking gun test, and the doubly decisive test), each of which provides different types of evidence (strong or weak, strengthening or weakening) for the mechanisms (see more detailed descriptions of these tests in appendix A, Glossary of Key Terms). In the Bayesian variant, confidence levels are often formalized in the fourth step.

Although process tracing is grounded in the qualitative tradition, the method is entirely compatible with and in many situations improved by the use of a broad range of both qualitative and quantitative data, including documents, media transcripts, interview or focus group transcripts, and findings from surveys. The type of data to be collected can sometimes appear unusual in comparison with other approaches, being akin to the type of evidence that may be produced in a court of law. The (formal or informal) Bayesian logic of the approach is indeed applied whenever the evidence must be rigorously evaluated to serve some probative function, and mysterious realities are to be understood or uncovered: in medical diagnosis, crime investigation, and courts of law.

THE ADVANTAGES AND DISADVANTAGES OF THE APPROACH

Process tracing serves well to develop and test hypotheses about the underlying processes generating program outcomes. As such, the method is particularly relevant when the aim of the evaluation is to explain how and under what circumstances a program works (or fails to work). A key strength of process tracing is that (especially in its Bayesian variant) it relies on a rigorous testing process where the probative value of data for given theories is established in a structured, transparent, and replicable manner. It thus has an advantage over similar approaches, particularly when theories are broad, vague, and ambiguous: the painstaking assessment of empirical data that the approach requires forces the evaluator to increase the precision and specificity of claims. By explicitly focusing on the weight of evidence in favor of or against specific mechanisms, the approach enhances the credibility of the conclusions drawn.

One limitation of process tracing is the difficulty of deciding how deep, how wide, and how far back to trace the causal processes. This limitation holds practical implications because the depth and breadth of the hypotheses considered determines the amount of data to be collected and analyzed (for best practice recommendations on this challenge, see Bennett and Checkel [2015] in Readings and Resources). This challenge can be mitigated with Bayesian formalization and by establishing an explicit confidence level at which the analysis is considered complete. Another challenge relates to the single-case nature of the analysis, which limits the generalizability of the findings. Finally, a practical obstacle is that the approach's thirst for "forensic proof" may be met with resistance from stakeholders; for example, documentation that would prove the existence of important parts of processes or mechanisms might be withheld for lack of trust or because sharing it is considered a breach of confidentiality.

THE APPLICABILITY OF THE APPROACH

As a single–case study approach, the process tracing approach may be widely applicable in development evaluation in general and IEO evaluations specifically. Published applications of process tracing in development evaluation are few but growing:

1. Process tracing was used to understand how citizen engagement affected intended outcomes in selected World Bank projects.

   (*Source*: World Bank. 2018. *Engaging Citizens for Better Development Results*. Independent Evaluation Group. Washington, DC: World Bank. https://ieg.worldbankgroup.org/evaluations/engaging-citizens-better-development-results.)

2. Process tracing was used to examine the effects of local civil society–led gender-responsive budgeting on maternal health service delivery in Kabale district in rural Uganda.

   (*Source*: Bamanyaki, P. A., and N. Holvoet. 2016. “Integrating Theory-Based Evaluation and Process Tracing in the Evaluation of Civil Society Gender Budget Initiatives.” *Evaluation* 22 (1): 72–90.)

3. Process tracing was used to understand how the influence process unfolded in an evaluation of the policy impact of the Uganda Poverty Conservation and Learning Group.

   (*Source*: D’Errico, S., B. Befani, F. Booker, and A. Giuliani. 2017. *Influencing Policy Change in Uganda: An Impact Evaluation of the Uganda Poverty and Conservation Learning Group’s Work*. London: International Institute for Environment and Development. http://pubs.iied.org/G04157/.)

4. Process tracing was used to evaluate the governance effectiveness of budget support interventions.

   (*Source*: Schmitt, J., and D. Beach. 2015. “The Contribution of Process Tracing to Theory-Based Evaluations of Complex Aid Instruments.” *Evaluation* 21 (4): 429–47.)

5. Process tracing was used to evaluate the policy impact of the Hunger and Nutrition Commitment Index.

   (*Source*: te Lintelo, D. J. H., T. Munslow, K. Pittore, and R. Lakshman. 2019. “Process Tracing the Policy Impact of ‘Indicators.’” *European Journal of Development Research* 32: 1338.)

6. Process tracing was used to assess a program’s contribution to reducing the number of minors working in the adult entertainment sector in Nepal.

(*Source*: Progress, Inc. and The Freedom Fund. 2020. *Evaluation of the Central Nepal Hotspot Project Using the Process Tracing Methodology—Summary Report*. London and New York: The Freedom Fund. https://freedomfund.org/our-reports/evaluation-of-the-central-nepal-hotspot-project-using-the-process-tracing-methodology/.)

7. Process tracing was used to understand the mechanisms (in addition to price reduction) through which the sugar-sweetened beverage tax worked in Barbados.

   (*Source*: Alvarado, M., T. Penney, and J. Adams. 2019. "OP113 Seeking Causal Explanations in Policy Evaluation: An Assessment of Applying Process Tracing to the Barbados Sugar-Sweetened Beverage Tax Evaluation." *Journal of Epidemiology and Community Health* 73: A53–A54.)

8. Process tracing was used to test the so-called kaleidoscope model of policy change. A set of 16 operational hypotheses identified the conditions under which food security interventions emerged on the policy agenda and were implemented in Zambia.

   (*Source*: Resnick, Danielle, Steven Haggblade, Suresh Babu, Sheryl L. Hendriks, and David Mather. 2018. "The Kaleidoscope Model of Policy Change: Applications to Food Security Policy in Zambia." *World Development* 109: 101–20.)

READINGS AND RESOURCES

Background

Beach, D., and R. Pedersen. 2013. *Process-Tracing Methods: Foundations and Guidelines*. Ann Arbor, MI: University of Michigan Press.

Befani, B., S. D'Errico, F. Booker, and A. Giuliani. 2016. "Clearing the Fog: New Tools for Improving the Credibility of Impact Claims." IIED Briefing. London: International Institute for Environment and Development. http://pubs.iied.org/17359IIED/.

Bennett, Andrew. 2010. "Process Tracing and Causal Inference." In *Rethinking Social Inquiry*, edited by Henry Brady and David Collier, chapter 10. Lanham, MD: Rowman & Littlefield. http://philsci-archive.pitt.edu/8872/.

Bennett, A., and J. T. Checkel. 2015. *Process Tracing: From Metaphor to Analytical Tool*. Cambridge, UK: Cambridge University Press.

Checkel, Jeffrey T. 2006. "Tracing Causal Mechanisms." *International Studies Review* 8 (2): 362–70.

Collier, D. 2011. "Understanding Process Tracing." *PS: Political Science & Politics* 44 (4): 823–30. https://polisci.berkeley.edu/sites/default/files/people/u3827/Understanding%20Process%20Tracing.pdf; https://www.cambridge.org/core/journals/ps-political-science-and-politics/article/understanding-process-tracing/183A057AD6A36783E678CB37440346D1/core-reader.

Kincaid, H., and D. Waldner. 2012. "Process Tracing and Causal Mechanisms." In *The Oxford Handbook of Philosophy of Social Science*. Oxford: Oxford University Press. https://www.oxfordhandbooks.com/view/10.1093/oxfordhb/9780195392753.001.0001/oxfordhb-9780195392753-e-4.

Palier, Bruno, and Christine Trampusch. 2016. "Process Tracing: The Understandings of Causal Mechanisms." Special issue, *New Political Economy* 21 (5). https://www.tandfonline.com/toc/cnpe20/21/5.

Ricks, J., and A. Liu. 2018. "Process-Tracing Research Designs: A Practical Guide." *PS: Political Science & Politics* 51 (4): 842–46.

Vennesson, P. 2008. "Case Studies and Process Tracing: Theories and Practices." In *Approaches and Methodologies in the Social Sciences: A Pluralist Perspective*, edited by D. Della Porta and M. Keating, 223–39. Cambridge: Cambridge University Press. https://minorthesis.files.wordpress.com/2012/12/vennesson-case-study-methods.pdf.

Advanced

Befani, B. 2016. "Testing Contribution Claims with Bayesian Updating." Evaluation Policy Practice Note 2.1 (Winter), CECAN, London. https://drive.google.com/file/d/0B-yPZxBK8WjPdEJNNUx5SG1VSFk/view.

Befani, B. 2020. "Quality of Quality: A Diagnostic Approach to Qualitative Evaluation." *Evaluation*.

Befani, B. Forthcoming. "Diagnostic Evaluation and Bayesian Updating: Practical Solutions to Common Problems." *Evaluation*.

Befani, B., and J. Mayne. 2014. "Process Tracing and Contribution Analysis: A Combined Approach to Generative Causal Inference for Impact Evaluation." *IDS Bulletin XLV* (6): 17–36.

Befani, B., and G. Steadman-Bryce. 2017. "Process Tracing and Bayesian Updating for Impact Evaluation." *Evaluation* 23 (1): 42–60.

Bennett, Andrew. 2008. "Process Tracing: A Bayesian Perspective." In *The Oxford Handbook of Political Methodology*, edited by Janet M. Box-Steffensmeier, Henry E. Brady, and David Collier. Oxford: Oxford University Press.

Elman, Colin, and John M. Owen, eds. 2015. “Process Tracing: A Symposium.” Special issue, *Security Studies* 24 (2). https://www.tandfonline.com/toc/fsst20/24/2?nav=tocList.

Fairfield, Tasha, and Andrew Charman. 2015. *Formal Bayesian Process Tracing: Guidelines, Opportunities, and Caveats.* London: The London School of Economics and Political Science. http://eprints.lse.ac.uk/62368/.

Fairfield, T., and A. Charman. 2017. “Explicit Bayesian Analysis for Process Tracing: Guidelines, Opportunities, and Caveats.” *Political Analysis* 25 (3): 363–80.

Tansey, O. 2007. “Process Tracing and Elite Interviewing: A Case for Non-probability Sampling.” *PS: Political Science & Politics* 40 (4): 765–72.

Other Resources

Befani, B. 2017. “Dr. Barbara Befani’s Bayes Formula Confidence Updater Spreadsheet.” Toolkits. CECAN, London. http://cecan.ac.uk/resources. This Excel-based template provides easy-to-use worksheets for updating confidence estimates when using Bayes updating.

6 Qualitative Comparative Analysis

BRIEF DESCRIPTION OF THE APPROACH

Qualitative comparative analysis (QCA) is a case-based analytical approach for identifying the causal conditions (for example, contexts or specific program components) that either individually or collectively generate a specific outcome. Its primary purpose is to identify and describe these causal conditions across a set of cases. This type of information is relevant to understanding and explaining how programs generate (or fail to generate) desired outcomes. What differentiates this approach from most other cross-case comparative methods is that it provides a specific set of algorithms to analyze data sets (usually in the form of a table). In this sense QCA can also be considered a data analysis technique. QCA is traditionally applied in cross-case analysis of multiple–case study designs.

Evaluation questions that QCA may answer include the following:

1. Under what circumstances did the program generate or not generate the desired outcome?
2. Which program components or ingredients (individually or collectively) are necessary or sufficient for the program outcome?
3. Are there alternative explanations for the desired outcome (or lack thereof)?

THE MAIN VARIATIONS OF THE APPROACH

In broad terms, there are two main variations of QCA: *crisp-set QCA* and *fuzzy-set QCA*. The distinction between them concerns how the causal conditions and outcomes are coded in the cases being studied.

In the traditional crisp-set QCA, each causal condition (and the outcome of interest) is coded as either present (1) or absent (0) in the cases to be analyzed. To illustrate, the use of a specific curriculum may be either present or absent in a given after-school program.

Fuzzy-set QCA uses more nuanced coding. It allows for causal conditions to be present or absent by degree. For instance, a specific causal condition may be fully implemented (coded 1), almost fully implemented (0.9), barely implemented (0.1), or not at all implemented (0). In the example of the after-school program, the use of the curriculum across schools may likely vary by degree among the teachers, with some teachers using it often (full implementation) and others using it only occasionally (barely implemented).

THE MAIN PROCEDURAL STEPS OF THE APPROACH

The application of QCA consists of seven steps:

- Identifying and defining the causal conditions and outcomes of interest;
- Assembling relevant data on each case included in the analysis (as informed by the causal conditions and outcomes of interest);
- Coding each case according to the presence or absence (dichotomously or by degree) of each causal condition and outcome;
- Using QCA software to summarize all the different causal configurations present among the cases;
- Using QCA software to simplify the identified configurations into the essential set of causal recipes eliciting a positive or negative outcome;
- Examining the consistency and empirical coverage of these recipes; and
- Reexamining the individual cases represented by each of the identified causal recipes to better understand the nature of the latter.

In practice, the steps are often applied iteratively, with the evaluator testing different causal models, perhaps moving back and forth between examination of the causal recipes identified and refining the way the cases have been coded.

Numerous software packages, some free of charge, are available for both crisp-set and fuzzy-set QCA (see Other Resources).

THE ADVANTAGES AND DISADVANTAGES OF THE APPROACH

Benefits of QCA include the ability to handle causal complexity (including conflicting cases), to identify different combinations of necessary and sufficient conditions associated with the same outcome, and to help in explaining how the outcome is generated across a small, medium, or large set of cases. From an internal validity perspective, a noteworthy strength of QCA is that the formalization of the logical comparisons (using Boolean algebra) provides for a systematic, transparent, and fully replicable analysis of qualitative data. Moreover, synthesizing medium or large data sets with QCA software allows for identification of general patterns in the data that would be impossible to capture manually. Finally, despite its generalization capabilities, QCA can be used with relatively small samples.

One possible limitation of QCA is that the method is difficult to use with a large number of causal conditions, especially when the set of available cases is small. Finding a good balance between consistency (the sufficiency of each pathway), coverage (the number of cases represented by the pathways), and parsimony or a manageable level of complexity of the findings usually requires that an experienced analyst work with the evaluation team on successive iterations and test increasingly simple models while maintaining high coverage and consistency. Theory often plays a key role in the initial selection, but confirmation bias is avoided because the results can reject those assumptions very strongly if the theory is not supported empirically. Therefore multiple iterations are needed until a set of pathways is found that reaches the optimal balance among consistency, coverage, and complexity.

Two additional, perhaps more practical, challenges relate to (i) the need for highly comparable data across all the cases to start the analysis; and (ii) the amount of work needed for full transparency on how the cases are coded (especially when fuzzy sets are used and when the coding has not been highly structured) and to systematically compare the information available for each condition across all cases. The coding must be systematically traceable to case-level information for the analysis to be fully replicable.

THE APPLICABILITY OF THE APPROACH

The introduction of QCA in development evaluations is fairly recent. Grounded on multiple–case study design, the QCA approach is widely applicable in both development evaluation in general and IEO evaluations specifically (the only requirement being the availability of comparable information across all cases). Examples include the following:

1. QCA was used to understand which factors contributed to a series of outcomes in carbon reduction interventions.

   (*Source*: World Bank. 2018. *Carbon Markets for Greenhouse Gas Emission Reduction in a Warming World*. Independent Evaluation Group. Washington, DC: World Bank. https://ieg.worldbankgroup.org/evaluations/carbon-finance.)

2. QCA was used in the impact evaluation of the Global Environment Facility / United Nations Development Programme Biodiversity, Protected Areas, and Protected Area Systems program.

   (*Source*: Befani, B. 2016. *Pathways to Change: Evaluating Development Interventions with Qualitative Comparative Analysis (QCA)*. Report 05/16, EBA, Stockholm. http://eba.se/wp-content/uploads/2016/07/QCA_BarbaraBefani-201605.pdf.)

3. QCA was applied in an evaluation of the effectiveness of gender-sensitive budget support in education.

   (*Source*: Holvoet, N., and L. Inberg. 2013. "Multiple Pathways to Gender-Sensitive Budget Support in the Education Sector." Working Paper 105, United Nations University World Institute for Development Economics Research, Helsinki.)

4. QCA was applied to better understand how and why Omar's Dream—a program that aims to end child labor—worked.

   (*Source*: Millard, A., A. Basu, K. Forss, B. Kandyomunda, C. McEvoy, and A. Woldeyohannes. 2015. *Is the End of Child Labour in Sight? A Critical Review of a Vision and Journey*. Geneva: International Cocoa Initiative. https://cocoainitiative.org/wp-content/uploads/2017/10/8_HIVOS.pdf.)

5. QCA was applied to identify factors affecting success in rendering water services sustainable.

   (*Source*: Welle, K., J. Williams, J. Pearce, and B. Befani. 2015. *Testing the Waters: A Qualitative Comparative Analysis of the Factors Affecting Success in Rendering Water Services Sustainable Based on ICT Reporting*. Brighton, UK: Institute of Development Studies and WaterAid. http://itad.com/wp-content/uploads/2015/09/MAVC_WaterAid_FINAL-report.pdf.)

6. QCA was applied in a macro evaluation of 50 UK Department for International Development social accountability projects to better understand what works, for whom, in what contexts, and why.

   (*Source*: Holland, J., F. Schatz, B. Befani, and C. Hughes. 2016. *Macro Evaluation of DFID's Policy Frame for Empowerment and Accountability*. Brighton, UK, and Washington, DC: ITAD. https://itad.com/wp-content/uploads/2017/06/EA-Macro-Evaluation-Technical-report-Dec16-FINAL.pdf.)

7. QCA was applied in an evaluation of a development cooperation program.

   (*Source*: Pattyn, V., A. Molenveld, and B. Befani. 2019. “Qualitative Comparative Analysis as an Evaluation Tool: Lessons from an Application in Development Cooperation.” *American Journal of Evaluation* 40 (1): 55–74.)

8. QCA was used to learn lessons from community forest management, comparing and synthesizing ten such cases from Africa, Asia, and Latin America.

   (*Source*: Arts, Bas, and Jessica de Koning. 2017. “Community Forest Management: An Assessment and Explanation of Its Performance through QCA.” *World Developmen*t 96: 315–25.)

9. QCA was used to synthesize fragmented studies, accumulate knowledge, and develop theory in water resource management.

   (*Source*: Mollinga, P., and D. Gondhalekar. 2014. “Finding Structure in Diversity: A Stepwise Small-N/Medium-N Qualitative Comparative Analysis Approach for Water Resources Management Research.” *Water Alternatives* 7 (1): 178–98. https://core.ac.uk/download/pdf/42549267.pdf.)

10. QCA has been recommended as a useful tool to understand the complex two-way relationship between migration and development.

    (*Source*: Czaika, Mathias, and Marie Godin. 2019. “Qualitative Comparative Analysis for Migration and Development Research.” MIGNEX Background Paper. Oslo: Peace Research Institute Oslo www.mignex.org/d022.)

11. QCA was used to understand the reasons behind development research uptake.

    (*Source*: Scholz, Vera, Amy Kirbyshire, and Nigel Simister. 2016. “Shedding Light on Causal Recipes for Development Research Uptake: Applying Qualitative Comparative Analysis to Understand Reasons for Research Uptake.” London: INTRAC and CKDN. https://www.intrac.org/resources/shedding-light-causal-recipes-development-research-uptake-applying-qualitative-comparative-analysis-understand-reasons-research-uptake/.)

12. QCA was used to understand why programs aimed at monitoring water quality succeed or fail.

(*Source*: Peletz, Rachel, Joyce Kisiangani, Mateyo Bonham, Patrick Ronoh, Caroline Delaire, Emily Kumpel, Sara Marks, and Ranjiv Khush. 2018. "Why Do Water Quality Monitoring Programs Succeed or Fail? A Qualitative Comparative Analysis of Regulated Testing Systems in Sub-Saharan Africa." *International Journal of Hygiene and Environmental Health* 221 (6): 907–20.)

READINGS AND RESOURCES

Background

Baptist, C., and B. Befani. 2015. "Qualitative Comparative Analysis—A Rigorous Qualitative Method for Assessing Impact." Coffey How To Paper. Better Evaluation, Melbourne, Victoria. https://www.betterevaluation.org/sites/default/files/Qualitative-Comparative-Analysis-June-2015%20(1).

Befani, B. 2016. *Pathways to Change: Evaluating Development Interventions with Qualitative Comparative Analysis (QCA)*. Report 05/16, EBA, Stockholm. http://eba.se/wp-content/uploads/2016/07/QCA_BarbaraBefani-201605.pdf.

Byrne, D. 2016. "Qualitative Comparative Analysis: A Pragmatic Method for Evaluating Intervention." Evaluation and Policy Practice Note 1 (Autumn), CECAN, London. https://www.cecan.ac.uk/sites/default/files/2018-01/DAVE%20B%20PPN%20v2.1.pdf.

Kahwati, L. C., and H. L. Kane. 2018. *Qualitative Comparative Analysis in Mixed Methods Research and Evaluation*. Thousand Oaks, CA: SAGE.

Ragin, Charles. 2014. *The Comparative Method: Moving beyond Qualitative and Quantitative Strategies*. Berkeley: University of California Press.

Rihoux, Benoît, and Bojana Lobe. 2009. "The Case for Qualitative Comparative Analysis (QCA): Adding Leverage for Thick Cross-Case Comparison." In *The SAGE Handbook of Case-Based Methods*, edited by D. S. Byrne and Charles C. Ragin, 222–42. Thousand Oaks, CA: SAGE.

Rihoux, Benoît, and Charles Ragin, eds. 2008. *Configurational Comparative Methods: Qualitative Comparative Analysis (QCA) and Related Techniques*. Thousand Oaks, CA: SAGE.

Schneider, C. Q., and C. Wagemann. 2012. *Set-Theoretic Methods for the Social Sciences—A Guide to Qualitative Comparative Analysis*. New York: Cambridge University Press.

Advanced

Befani, B., S. Ledermann, and F. Sager. 2007. “Realistic Evaluation and QCA: Conceptual Parallels and an Empirical Application.” *Evaluation* 13 (2): 171–92.

Blackman, T., J. Wistow, and D. Byrne. 2013. “Using Qualitative Comparative Analysis to Understand Complex Policy Problems.” *Evaluation* 19 (2): 126–40.

Duşa, Adrian. 2019. *QCA with R: A Comprehensive Resource*. Cham, Switzerland: Springer.

Grofman, B., and C. Q. Schneider. 2009. “An Introduction to Crisp Set QCA, with a Comparison to Binary Logistic Regression.” *Political Research Quarterly* 62 (4): 662–72.

Hudson, John, and Stefan Kühner, eds. 2013. “Innovative Methods for Policy Analysis: QCA and Fuzzy Sets.” Special issue, *Policy and Society* 32 (4).

Marx, A., and A. Duşa. 2011. “Crisp-Set Qualitative Comparative Analysis (csQCA), Contradictions and Consistency Benchmarks for Model Specification.” *Methodological Innovations Online* 6 (2): 103–48.

Ragin, C. C., D. Shulman, A. Weinberg, and B. Gran. 2003. “Complexity, Generality, and Qualitative Comparative Analysis.” *Field Methods* 15 (4): 323–40.

Sager, F., and C. Andereggen. 2012. “Dealing with Complex Causality in Realist Synthesis: The Promise of Qualitative Comparative Analysis.” *American Journal of Evaluation* 33 (1): 60–78.

Thomann, E., and M. Maggetti. 2020. “Designing Research with Qualitative Comparative Analysis (QCA): Approaches, Challenges, and Tools.” *Sociological Methods & Research* 49 (2): 356–86.

Thomas, James, Alison O’Mara-Eves, and Ginny Brunton. 2014. “Using Qualitative Comparative Analysis (QCA) in Systematic Reviews of Complex Interventions: A Worked Example.” *Systematic Reviews* 3, article 67.

Verweij, S., and L. M. Gerrits. 2013. “Understanding and Researching Complexity with Qualitative Comparative Analysis: Evaluating Transportation Infrastructure Projects.” *Evaluation* 19 (1): 40–55.

Other Resources

COMPASSS (COMPArative Methods for Systematic cross-caSe analySis) is a worldwide network bringing together scholars and practitioners interested in the further development and application of configurational comparative and set-theoretical methods (crisp-set QCA, multivalue QCA, fuzzy-set QCA, and linked methods and techniques). www.compasss.org.

Ragin, C. C., K. A. Drass, and S. Davey. 2006. “Fuzzy-Set/Qualitative Comparative Analysis 2.0.” Department of Sociology, University of Arizona, Tucson. http://www.u.arizona.edu/~cragin/fsQCA/software.shtml. This free online software facilitates crisp- and fuzzy-set QCA, including development of truth tables and complex, intermediate, and parsimonious solutions. A user manual supports the application of the workbooks.

7 Participatory Evaluation

BRIEF DESCRIPTION OF THE APPROACH

Participatory approaches emphasize stakeholder involvement (program staff and beneficiaries, among others) in all or most of the design, implementation, and reporting stages of an evaluation. Participatory evaluation is often motivated by the idea that actively involving stakeholders (including those affected by the program) in the evaluation process gives them a voice in how the evaluation is designed and implemented, promotes a sense of ownership and empowerment, and enhances the potential relevance and use of the evaluation. Participatory approaches tend to be more relevant when the primary purpose is to provide information for program improvement or organizational development and not necessarily to make definitive statements about program outcomes.

Evaluation questions that participatory approaches may answer include the following:

1. What are the primary goals or outcomes of the program from the perspective of different stakeholders?
2. What program services are needed? How are they best delivered?
3. How was the program implemented?

THE MAIN VARIATIONS OF THE APPROACH

Participatory approaches can be applied in combination with any other evaluation approach or method. Its many variations reach far beyond the scope of this guidance note. A distinction has been made between *pragmatic* and *transformative participatory approaches*. Pragmatic approaches are motivated by the practical benefits of including stakeholders (including increased use of findings), and transformative approaches aim to change specific conditions for the stakeholders. One of the most widely used participatory approaches is *utilization-focused evaluation* (see Readings and Resources for more detailed information on this approach).

THE MAIN PROCEDURAL STEPS OF THE APPROACH

Given the diversity of approaches and methods, a common set of procedural steps cannot be identified for participatory approaches to evaluation. However, a central step in all participatory approaches is carefully considering and defining the stakeholders to be included (for example, those who benefit from the program, those who influence the implementation of the program, those who oppose the program), the scope and nature of their involvement, and their concrete role and responsibilities in the evaluation. The selection of which stakeholders to include (and, in effect, which not to include) is a core step in any participatory approach, and it holds significant methodological and ethical implications.

THE ADVANTAGES AND DISADVANTAGES OF THE APPROACH

Some potential benefits of participatory approaches include enhancing the cultural responsiveness and relevance of the evaluation, building capacity and empowering local stakeholders, and enhancing the authenticity and accuracy of the data and findings. To realize these benefits, though, careful consideration should be given to the stakeholders to be included (whose voices are included?); their intended role and purpose in the evaluation (are they identifying relevant questions, or are they consulted in the collection and analysis of data?); the extent of their expected participation in the evaluation (what is the breadth and depth of the engagement?); how they will be involved in the different stages of the evaluation (what is the format of and process for their involvement?); the ability and capacity of the stakeholders to actively engage and participate (is skill development or training called for?); and the value of participation for the stakeholders.

Common concerns for participatory evaluation include the potential burden and cost of participation among the stakeholders (especially when these outweigh the associated benefits) and the difficulty of engaging stakeholders without reinforcing existing power hierarchies (participation might be biased toward specific groups of stakeholders). In multilevel, multisite evaluations, stakeholder participation must be weighed carefully against the costs and the potential benefits of including stakeholders from across the (complex) program.

THE APPLICABILITY OF THE APPROACH

Participatory approaches are highly applicable in development evaluation in general and for certain types of evaluation. IEOs generally would not have direct involvement with the intervention or its stakeholders. In practice, this means that several variations of participatory evaluation are not applicable. Nonetheless, creative applications of participatory approaches that safeguard independence and the efficient incorporation of stakeholder inputs into IEO evaluations are encouraged in most IEOs today. The implementation of participatory approaches requires in-depth knowledge of the context of the stakeholders (especially in relation to the program) and facilitation skills to manage stakeholder interactions (see also guidance note 13, Focus Group).

The practical applications of participatory approaches cover a broad range of programs and sectors:

1. Participatory assessment was used as part of an evaluation of the effectiveness of a water and sanitation program in Flores, Indonesia. Marginalized groups, women, and the poor were included through local gender-balanced teams of evaluators.

   (*Source*: Sijbesma, Christine, Kumala Sari, Nina Shatifan, Ruth Walujan, Ishani Mukherjee, and Richard Hopkins. 2002. *Flores Revisited: Sustainability, Hygiene and Use of Community-Managed Water Supply and Sanitation and the Relationships with Project Approaches and Rules*. Delft: IRC International Water and Sanitation Centre; Jakarta: WSP-EAP. https://www.ircwash.org/sites/default/files/Sijbesma-2002-Flores.pdf.)

2. Participatory assessment approaches have been used by the World Bank to assess beneficiaries' experiences and perceived benefits of agricultural programs in countries such as Ghana, Guinea, Mali, and Uganda.

   (*Source*: Salmen, L. F. 1999. "The Voice of the Farmer in Agricultural Extension. A Review of Beneficiary Assessment of Agricultural Extension and an Inquiry into Their Potential as a Management Tool." AKIS Discussion Paper, World Bank, Washington, DC. http://documents.worldbank.org/curated/en/776431468322742990/pdf/multi0page.pdf.)

3. In the Uganda Participatory Poverty Assessment Process, local people were consulted in 36 rural and urban sites in nine districts in Uganda. In this assessment, "voices" and perspectives of the poor were brought to

the fore to influence district and national planning, implementation, and monitoring.

(*Source*: Uganda Participatory Poverty Assessment Process. 2002. "Deepening the Understanding of Poverty—Second Participatory Poverty Assessment." Kampala: Uganda Ministry of Finance, Planning and Economic Development. http://www.participatorymethods.org/sites/participatorymethods.org/files/deepning%20the%20understanding%20of%20poverty.pdf.)

4. A participatory ethnographic evaluation and research approach was used in Cambodia and Myanmar to help inform program design by gaining an in-depth understanding of the sexual partners and clients of informal sex workers.

   (*Source*: World Bank. 2007. *Tools for Institutional, Political, and Social Analysis of Policy Reform. A Sourcebook for Development Practitioners*. Washington, DC: World Bank. https://siteresources.worldbank.org/EXTTOPPSISOU/Resources/1424002-1185304794278/TIPs_Sourcebook_English.pdf.)

5. Participatory wealth ranking was applied to almost 10,000 households to assess the number of poor households and their level of poverty in rural South Africa. Local perceptions of poverty were used to generate a wealth index of asset indicators.

   (*Source*: Hargreaves, J. R., L. A. Morison, J. S. S. Gear, M. B. Makhubele, J. Porter, J. Buzsa, C. Watts, J. C. Kim, and P. M. Pronk. 2007. "'Hearing the Voices of the Poor': Assigning Poverty Lines on the Basis of Local Perceptions of Poverty: A Quantitative Analysis of Qualitative Data from Participatory Wealth Ranking in Rural South Africa." *World Development* 35 (2): 212–19.)

6. A participatory approach was used in the development and application of a self-assessment of household livelihood viability index in Ethiopia. Qualitative case studies of livelihoods in select villages and group discussions with people from these villages were used in the development of the self-assessment tool.

   (*Source*: Chambers, R. 2007. "Who Counts? The Quiet Revolution of Participation and Numbers." Working Paper 296, IDS, Brighton. https://opendocs.ids.ac.uk/opendocs/handle/20.500.12413/398.)

READINGS AND RESOURCES

Background

Blackburn, James, and Jeremy Holland, eds. 1998. “Foreword.” In *Who Changes? Institutionalizing Participation in Development*. London: Intermediate Technology Publications. https://opendocs.ids.ac.uk/opendocs/bitstream/handle/20.500.12413/685/rc77.pdf.

Cousins, J. B., and E. Whitmore. 1998. “Framing Participatory Evaluation.” *New Directions for Evaluation* 80: 5–23.

Estrella, M., J. Blauert, J. Gonsalves, D. Campilan, J. Gaventa, I. Guijt, D. Johnson, and R. Ricafort, eds. 2000. *Learning from Change: Issues and Experiences in Participatory Monitoring and Evaluation*. London: Intermediate Technology Publications and International Development Research Center.

Estrella M., and J. Gaventa. 1998. “Who Counts Reality? Participatory Monitoring and Evaluation: A Literature Review.” Working Paper 70, IDS, Brighton, UK. https://opendocs.ids.ac.uk/opendocs/handle/20.500.12413/3388.

Pretty, Jules, Irene Guijt, John Thompson, and Ian Scoones. 1995. *Participatory Learning and Action: A Trainer’s Guide*. London: International Institute for Environment and Development. http://www.experience-capitalization.net/handle/123456789/60.

Whitmore, Elizabeth, ed. 1998. “Understanding and Practicing Participatory Evaluation.” Special issue, *New Directions for Evaluation* (80).

Advanced

Chambers, R. 2007. “Who Counts? The Quiet Revolution of Participation and Numbers.” Working Paper 296, IDS, Brighton, UK. https://opendocs.ids.ac.uk/opendocs/handle/20.500.12413/398.

Chouinard, J. A., and P. Milley. 2018. “Uncovering the Mysteries of Inclusion: Empirical and Methodological Possibilities in Participatory Evaluation in an International Context.” *Evaluation and Program Planning* 67: 70–78.

Cornwall, A. 2008. “Unpacking ‘Participation’: Models, Meanings and Practices.” *Community Development Journal* 43 (3): 269–83.

Cornwall, Andrea, and Alia Aghajanian. 2017. “How to Find Out What’s Really Going On: Understanding Impact through Participatory Process Evaluation.” *World Development* 99: 173–85.

Cornwall, A., and R. Jewkes. 1995. “What Is Participatory Research?” *Social Science & Medicine* 41 (12): 1667–76.

Dart, Jessica, and Rick Davies. 2003. “A Dialogical, Story-Based Evaluation Tool: The Most Significant Change Technique.” *American Journal of Evaluation* 24 (2): 137–55.

Guijt, I. 2000. “Methodological Issues in Participatory Monitoring and Evaluation.” In *Learning from Change: Issues and Experiences in Participatory Monitoring and Evaluation*, edited by M. Estrella, J. Blauert, D. Campilan, J. Gaventa, I. Guijt, D. Johnson, and R. Ricafort, 201–16. London: Intermediate Technology Publications.

Holland, Jeremy. 2013. *Who Counts? The Power of Participatory Statistics*. Rugby, UK: Practical Action.

Holland, Jeremy, and James Blackburn. 1998. *Whose Voice? Participatory Research and Policy Change*. London: Intermediate Technology Publications. https://opendocs.ids.ac.uk/opendocs/bitstream/handle/20.500.12413/686/rc78.pdf?sequence=1.

Lewis, David. 2018. “Peopling Policy Processes? Methodological Populism in the Bangladesh Health and Education Sectors.” *World Development* 108: 16–27.

Patton, M. Q. 2008. *Utilization-Focused Evaluation* (4th ed.). Thousand Oaks, CA: SAGE.

Pouw, N., T. Dietz, A. Bélemvire, D. de Groot, D. Millar, F. Obeng, W. Rijneveld, K. van der Geest, Z. Vlaminck, and F. Zaal. 2017. “Participatory Assessment of Development Interventions: Lessons Learned from a New Evaluation Methodology in Ghana and Burkina Faso.” *American Journal of Evaluation* 38 (1): 47–59.

Rossignoli, C. M., F. Di lacovo, R. Moruzzo, and P. Scarpellini. 2017. “Enhancing Participatory Evaluation in a Humanitarian Aid Project.” *Evaluation* 23 (2): 134–51.

Shulha, L. M., E. Whitmore, J. B. Cousins, N. Gilbert, and H. A. Hudib. 2016. “Introducing Evidence-Based Principles to Guide Collaborative Approaches to Evaluation: Results of an Empirical Process.” *American Journal of Evaluation* 37 (2): 193–215.

Other Resources

The Most Significant Change method is a form of participatory monitoring and evaluation, where many project stakeholders both choose the sorts of change to be recorded and analyze the data. It provides data on impact and outcomes that can be used to help assess the overall performance of a program. https://www.odi.org/publications/5211-strategy-development-most-significant-change-msc.

Participatory Learning and Action is a journal published by the International Institute for Environment and Development in collaboration with the Institute of Development Studies. https://www.iied.org/participatory-learning-action-pla.

The Participatory Methods website, managed by the Participation, Inclusion and Social Change Cluster at the Institute of Development Studies, provides resources to generate ideas and action for inclusive development and social change, including participatory approaches to program design, monitoring, and evaluation. https://www.participatorymethods.org/.

The US Agency for International Development's MEASURE Evaluation project has a stakeholder engagement tool that provides an organizing framework for identifying stakeholders; defining stakeholder roles and resources; assessing stakeholder interests, knowledge, and positions; creating an engagement plan; and tracking stakeholder engagement, among other things. https://www.measureevaluation.org/resources/publications/ms-11-46-e.

8 System Mapping and Dynamics

BRIEF DESCRIPTION OF THE APPROACH

System mapping and dynamics are visual approaches for better understanding the systemic nature of programs embedded in their contexts. The primary purpose of *system mapping* is to describe the different components of a system (microlevel content) and how these are connected (macrolevel structure). In logical extension, the purpose of *system dynamics* is to understand and describe how different microsystem components interact to generate macrolevel change (dynamic). In evaluation, system mapping and dynamics are particularly relevant for understanding, for example, the institutional, social, economic, and cultural aspects of the context in which a program operates and how they influence how the program works. This supports a better understanding of the nature and impact of complex programs.

Evaluation questions that system mapping and system dynamics may answer include the following:

1. How do program stakeholders interact among themselves and with their surroundings? How does this affect the program outcome?
2. How is the program affected by the wider, complex, systemic context it is embedded in? How does it adapt to its environment over time?

THE MAIN VARIATIONS OF THE APPROACH

System mapping may focus on a broad range of systems. To illustrate, system maps may be in the form of actor maps (describing how individuals or groups influencing a system are connected), mind or issue maps (describing trends and connections among different political, organizational, and cultural perspectives or issues), or causal loop diagrams (describing the causal pathways and dynamics within a system of interest), or even a combination of these.

System dynamics may vary in terms of the scope and level of detail aimed at in the description of the system, though it always entails breathing life into some sort of system map and making it dynamic. It requires the specification and use of system components interacting with each other. Some applications of system dynamics use causal loop diagrams as a complementary precursor to the development of stock-flow diagrams or Bayesian belief networks. Agent-based modeling focuses on interactions among agents occupying specific locations on a gridlike space.

THE MAIN PROCEDURAL STEPS OF THE APPROACH

System mapping involves four procedural steps:

- Defining the type of system and elements to be mapped;
- Specifying the boundaries of the map (what is within and outside the system of interest?);
- Often facilitating a workshop with stakeholders to identify the microlevel interacting elements that form the system and their behavior; and
- Finalizing the map.

Making system maps dynamic involves four major steps:

- Identifying the specific causal relationships to be modeled;
- Developing a working hypothesis of the microlevel causal relationships among system elements—these might be, for example, “stocks” (such as the number of vaccines) and “flows” (such as the rate of issuing vaccines, interest in vaccines) that collectively make up the macrolevel causal dynamic (such as increased poultry vaccinations)—or assigning characteristics and rules of behavior to agents;
- Empirically validating and refining the model by collecting data, presenting and discussing the model with relevant stakeholders, or both; and
- Presenting and using the refined model for future program planning and design.

THE ADVANTAGES AND DISADVANTAGES OF THE APPROACH

For evaluating programs characterized by significant causal complexity, system mapping is particularly valuable. The visual aspect of system mapping puts complex concepts and relationships into simpler pictorial representations. It becomes possible to visualize and describe nonlinear feedback loops, delays in outcomes, and unexpected collective outcomes that result from countless and complex microinteractions. This type of information supports program decision-making and design leading toward sustainable, systemic change.

System mapping and dynamics come with several challenges. The focus on defining and examining system boundaries is one of the most important and most difficult elements of the approaches. In real-world programs, it can often be difficult to define the boundaries of the system, and in effect the boundaries of the map or model, before the data collection. This happens in part because modeling real-world systems necessarily involves the imposition of a boundary that in some sense is arbitrary. In practice, then, the system boundaries may not solidify until data have been collected and examined. There is often confusion between either understanding complex relations or more faithfully representing an empirical reality. In addition, making systems dynamic requires technical expertise and the ability to use relevant software (for example, iThink, Vensim, and NetLogo).

THE APPLICABILITY OF THE APPROACH

Recently, there has been a surge of interest in systemic approaches to evaluation. System mapping is highly applicable to most evaluation settings, including IEOs, especially given the complexity of many of the interventions (sectors, themes, and so on) being evaluated. Many seasoned and newer evaluators would benefit from learning and employing this approach, but practical applications in development contexts are still relatively rare. Examples include the following:

1. System mapping was used in an evaluation for the humanitarian aid organization CARE. CARE engaged in a project to improve the organization's systems. System maps were used to help set data collection priorities and to guide data collection planning.

   (*Source*: Coffman, J., and E. Reed. 2009. "Unique Methods in Advocacy Evaluation." Methods Brief, Innovation Network, Washington, DC. http://www.pointk.org/resources/files/Unique_Methods_Brief.pdf.)

2. System dynamics was used in an evaluation of a malaria control program in Bolivia.

   (*Source*: Newman, J., M. A. Velasco, L. Martin, and A. M. Fantini. 2003. "A System Dynamics Approach to Monitoring and Evaluation at the Country Level: An Application to the Evaluation of Malaria-Control Programs in Bolivia." Paper presented at the Fifth Biennial World Bank Conference on Evaluation and Development, Washington, DC, 15—16 July. http://www.csdnet.aem.cornell.edu/papers/newman.pdf.)

3. A systems approach was applied in an evaluation of a peacebuilding initiative in Ghana, Guinea-Bissau, and Kosovo.

(*Source*: Chigas, D., and P. Woodrow. 2014. "Systems Thinking in Peace Building Evaluations: Applications in Ghana, Guinea-Bissau and Kosovo." In *Evaluation Methodologies for Aid in Conflict*, edited by O. W. Andersen, B. Bull, and M. Kennedy, 175–97. London, UK: Routledge.)

4. System dynamics modeling was used to compare the observed postprogram situation with a hypothetical nonintervention scenario as part of an impact evaluation of a private sector development program.

   (*Source*: Derwisch, S., and P. Löwe. 2015. "Systems Dynamics Modeling in Industrial Development Evaluation." *IDS Bulletin* 46 (1): 44–57.)

5. A system dynamics model was used to evaluate the potential requirements and implications on the health systems of the ambitious antiretroviral therapy scale-up strategy in Lusaka, Zambia.

   (*Source*: Grove, J. T. 2015. "Aiming for Utility in 'Systems-Based Evaluation': A Research-Based Framework for Practitioners." *IDS Bulletin* 46 (1): 58–70.)

READINGS AND RESOURCES

Background

Bamberger, M., J. Vaessen, and E. Raimondo. 2015. *Dealing with Complexity in Development Evaluation*. Thousand Oaks, CA: SAGE.

Fujita, N., ed. 2010. "Beyond Logframe: Using Systems Concepts in Evaluation." Issues and Prospects of Evaluations for International Development—Series IV. Tokyo: Foundation for Advanced Studies on International Development. https://www.perfeval.pol.ulaval.ca/sites/perfeval.pol.ulaval.ca/files/publication_129.pdf.

Hummelbrunner, R. 2011. "Systems Thinking and Evaluation." *Evaluation* 17 (4): 395–403.

Thomas, V. G., and B. A. Parsons. 2017. "Culturally Responsive Evaluation Meets Systems-Oriented Evaluation." *American Journal of Evaluation*, 38 (1): 7–28.

Williams, B. 2015. "Prosaic or Profound? The Adoption of Systems Ideas by Impact Evaluation." *IDS Bulletin* 46 (1): 7–16.

Williams, B., and R. Hummelbrunner. 2011. *Systems Concepts in Action: A Practitioner's Toolkit*. Stanford, CA: Stanford University Press.

Advanced

Barbrook-Johnson, Pete. 2020. “Participatory Systems Mapping in Action: Supporting the Evaluation of the Renewable Heat Incentive.’” Evaluation Policy and Practice Note 17 (April), CECAN, London. https://www.cecan.ac.uk/sites/default/files/2020-04/17%20Participatory%20Systems%20Mapping%20in%20action%20%28online%29.pdf.

Caffrey, L., and E. Munro. 2017. “A Systems Approach to Policy Evaluation.” *Evaluation* 23 (4): 463–78.

Canham, Sarah L., Mei Lan Fang, Lupin Battersby, and Mineko Wada. 2019. “Understanding the Functionality of Housing-Related Support Services through Mapping Methods and Dialogue.” *Evaluation and Program Planning* 72: 33–39.

Gates, E. F. 2018. “Toward Valuing with Critical Systems Heuristics.” *American Journal of Evaluation*, 39 (2): 201–20.

Gopal, S., and T. Clark. 2015. “System Mapping: A Guide to Developing Actor Maps.” The Intersector Project. http://intersector.com/resource/system-mapping-a-guide-to-developing-actor-maps/.

Guichard, Anne, Émilie Tardieu, Christian Dagenais, Kareen Nour, Ginette Lafontaine, and Valéry Ridde. 2017. “Use of Concurrent Mixed Methods Combining Concept Mapping and Focus Groups to Adapt a Health Equity Tool in Canada.” *Evaluation and Program Planning* 61: 169–77.

Lich, Kristen Hassmiller, Jennifer Brown Urban, Leah Frerichs, and Gaurav Dave. 2017. “Extending Systems Thinking in Planning and Evaluation Using Group Concept Mapping and System Dynamics to Tackle Complex Problems.” *Evaluation and Program Planning* 60: 254–64.

Sedighi, Tabassom. 2019. “A Bayesian Network for Policy Evaluation.” Evaluation Policy and Practice Note 13, CECAN, London. https://www.cecan.ac.uk/sites/default/files/2019-03/13%20Bayesian%20Network%20%28online%29.pdf.

Sterman, J. 2016. “Fine-Tuning Your Causal Loop Diagrams.” The Systems Thinker. https://thesystemsthinker.com/fine-tuning-your-causal-loop-diagrams-part-i/.

Trochim, William M., and Daniel McLinden. 2017. “Introduction to a Special Issue on Concept Mapping.” *Evaluation and Program Planning* 60: 166–75.

Uprichard, Emma, and Alexandra Penn. 2016. “Dependency Models.” Evaluation Policy and Practice Note 4, CECAN, London. https://www.cecan.ac.uk/sites/default/files/2018-01/EMMA%20PPN%20v1.0.pdf.

Vaughn, Lisa M., Jennifer R. Jones, Emily Booth, and Jessica G. Burke. 2017. “Concept Mapping Methodology and Community-Engaged Research: A Perfect Pairing.” Evaluation and Program Planning 60: 229–37.

Wilkinson, Helen, Nigel Gilbert, Liz Varga, Corinna Elsenbroich, Henry Leveson-Gower, and Jonathan Dennis. 2018. “Agent-Based Modelling for Evaluation.” *Evaluation Policy and Practice Note* 3, CECAN, London. https://www.cecan.ac.uk/sites/default/files/2018-01/HELEN%20ABM%20PPN%20v0.4.pdf.

9 Outcome Mapping and Outcome Harvesting

BRIEF DESCRIPTION OF THE APPROACH

Outcome mapping is a participatory approach for program planning, design, evaluation, and monitoring that centers on the outcomes of individuals, groups, or organizations with which the program has direct contact. *Outcome harvesting* is a qualitative approach that relies on testimonies from key program stakeholders to identify, formulate, and make sense of program outcomes (both intended and unintended).

Evaluation questions that outcome mapping and outcome harvesting may answer include the following:

1. What key program partners can influence the program implementation and performance?
2. What are the primary goals or outcomes of the program from the perspective of different stakeholders?
3. What did the program achieve that was either intended or unintended?
4. What, if any, program or stakeholder processes and activities contributed to the outcomes?

THE MAIN VARIATIONS OF THE APPROACH

Both outcome mapping and harvesting emphasize program learning and stakeholder collaboration. Whereas outcome mapping primarily focuses on the program or stakeholder processes that contribute to specific, prespecified outcomes, outcome harvesting focuses more deliberately on uncovering hidden or unexpected outcomes as identified by stakeholders. In marked contrast with outcome mapping, outcome harvesting does not focus on preestablished outcomes or progress markers, allowing instead the primary program outcomes to be identified and described ex post by the stakeholders.

Outcome mapping can be conducted prospectively or ex ante to identify what different stakeholders consider the intended or expected outcomes, while outcome harvesting is conducted retrospectively or ex post, with different stakeholders determining which outcomes have been realized.

THE MAIN PROCEDURAL STEPS OF THE APPROACH

Outcome mapping develops in three stages. In the design phase, the program vision or mission, intended changes (and change processes), and key stakeholders (referred to as boundary partners) are identified. The second step, program monitoring, involves ongoing self-assessment of the progress of the program toward the desired changes. The monitoring phase, aimed at learning and adaptation, is structured around specific progress markers and emphasizes the program strategy and activities contributing to these. In the third and final stage, the evaluation plan and priorities are developed to emphasize further program and stakeholder development. These three stages have been formalized into 12 specific operational steps.

Outcome harvesting typically involves the following six steps:

- Designing the harvest, which involves deciding on the questions to be addressed, which then inform decisions on whom to collect outcome information from and the type of outcome information to be collected;
- Harvesting outcome testimonies through existing documents, surveys, interviews, or some combination of these;
- Formulating, sharing, and refining outcome stories with change agents (for example, program staff);
- Substantiating and validating the outcome stories with independent individuals who have knowledge of the outcomes;
- Analyzing the final set of revised outcome stories in relation to the initial questions motivating the outcome harvest; and
- Supporting the use of findings (for example, by connecting stories with decision-making on program refinements).

THE ADVANTAGES AND DISADVANTAGES OF THE APPROACH

Outcome mapping and harvesting are both appropriate for better understanding the changes that are occurring among stakeholders with immediate, direct program contact. They also are useful when the emphasis is on stakeholder collaboration and capacity building and on the relative contribution of the program, that is, the program's effects in combination with other causal factors (as opposed to its quantifiable net effect). Although outcome mapping (as a prospective approach) is most effective for the planning stages of an evaluation, outcome harvesting (as a retroactive approach) reaches beyond planning and into the implementation of the evaluation. Both approaches also serve well to identify unintended consequences of a program and other contributing factors to observed outcomes.

Limitations to outcome mapping and harvesting include the relatively light examination of the wider outcomes and influences of the program. This trade-off is necessary given the explicit focus on stakeholders with direct program contact and the parallel emphasis on immediate program outcomes. The causal inference aspects are also scientifically weak, with no causal model being explicitly used to assess contributions to outcomes. In addition, the connection between empirical observations and theories is weak: it is unclear how data become evidence of an outcome and how strongly they support theories. Another more practical challenge is the collaborative inclusion of stakeholders, which often requires resources and time for training and facilitation. As in participatory evaluation more broadly, this makes outcome mapping and outcome harvesting more difficult to use in complex (multilevel, multisite) programs.

THE APPLICABILITY OF THE APPROACH

Both outcome mapping and harvesting are widely used in development evaluation in general. A practical difficulty in their employment by IEOs, however, is their necessarily participatory nature. As such, outcome mapping and harvesting are often used in development evaluation by teams directly involved in (often relatively small-scale) interventions or directly working with project teams to provide outcome mapping services (such as evaluators providing advisory services).

Examples include the following:

1. Outcome mapping was used in a women's health and empowerment program in India to document and assess women's own capacity development in gender issues, monitoring and evaluation, and applied research.

   (*Source*: Earl, S., F. Carden, and F. Smutylo. 2001. *Outcome Mapping: Building Learning and Reflection into Development Programs*. Ottawa: International Development Research Centre. https://www.idrc.ca/en/book/outcome-mapping-building-learning-and-reflection-development-programs.)

2. Outcome mapping was used to identify changes in stakeholder attitudes toward the forest and their agricultural land as part of an evaluation of the Caja Andina Project in Ecuador.

(*Source*: Smutylo, T. 2005. “Outcome Mapping: A Method for Tracking Behavioral Changes in Development Programs.” ILAC Brief 7 (August), Consultative Group on International Agricultural Research, Montpellier, France. https://cgspace.cgiar.org/bitstream/handle/10568/70174/ILAC_Brief07_mapping.pdf?sequence=1&isAllowed=y.)

3. Outcome mapping was used to develop a monitoring framework for agricultural market development projects in Bolivia, Ecuador, and Peru.

   (*Source*: Smutylo, T. 2005. “Outcome Mapping: A Method for Tracking Behavioral Changes in Development Programs.” ILAC Brief 7 (August), Consultative Group on International Agricultural Research, Montpellier, France. https://cgspace.cgiar.org/bitstream/handle/10568/70174/ILAC_Brief07_mapping.pdf?sequence=1&isAllowed=y.)

4. Outcome harvesting was used to identify and analyze unintended outcomes arising from the Alliances Lesser Caucasus Programme’s activities in the dairy industry in Kvemo Kartli, Georgia.

   (*Source*: USAID [US Agency for International Development]. 2016. “Testing Tools for Assessing Systemic Change: Outcome Harvesting—The ALCP Project in the Georgian Dairy Industry.” LEO Report 43, USAID, Washington, DC. https://www.marketlinks.org/sites/marketlinks.org/files/resource/files/Report_No._43_-_SC_Tool_Trial_Outcome_Harvesting_-_508_compliant3.pdf.)

5. Outcome harvesting was used to improve governance in pharmaceutical procurement and supply chain management in Kenya, Tanzania, and Uganda.

   (*Source*: World Bank. 2014. *Cases in Outcome Harvesting: Ten Pilot Experiences Identify New Learning from Multi-Stakeholder Projects to Improve Results*. Washington, DC: World Bank. https://openknowledge.worldbank.org/handle/10986/20015.)

6. Outcome harvesting was used in an evaluation of a program developing capacity and delivering results in public sector reform in Burundi.

   (*Source*: World Bank. 2014. *Cases in Outcome Harvesting: Ten Pilot Experiences Identify New Learning from Multi-Stakeholder Projects to*

Improve Results. Washington, DC: World Bank. https://openknowledge.worldbank.org/handle/10986/20015.)

7. Outcome harvesting was used in an evaluation of a program strengthening parliamentary oversight of national budgets in Africa.

   (*Source*: World Bank. 2014. *Cases in Outcome Harvesting: Ten Pilot Experiences Identify New Learning from Multi-Stakeholder Projects to Improve Results*. Washington, DC: World Bank. https://openknowledge.worldbank.org/handle/10986/20015.)

8. Outcome harvesting was used in an evaluation of a program increasing capacity development of city officials and practitioners across China through e-learning.

   (*Source*: World Bank. 2014. *Cases in Outcome Harvesting: Ten Pilot Experiences Identify New Learning from Multi-Stakeholder Projects to Improve Results*. Washington, DC: World Bank. https://openknowledge.worldbank.org/handle/10986/20015.)

9. Outcome harvesting was used in an evaluation of a program strengthening implementation of legislation on access to information across Latin America.

   (*Source*: World Bank. 2014. *Cases in Outcome Harvesting: Ten Pilot Experiences Identify New Learning from Multi-Stakeholder Projects to Improve Results*. Washington, DC: World Bank. https://openknowledge.worldbank.org/handle/10986/20015.)

READINGS AND RESOURCES

Background

Jones, H., and S. Hearn. 2009. "Outcome Mapping: A Realistic Alternative For Planning, Monitoring, And Evaluation." Background Note (October), Overseas Development Institute, London. https://www.odi.org/sites/odi.org.uk/files/odi-assets/publications-opinion-files/5058.pdf.

Smutylo, T. 2005. "Outcome Mapping: A Method for Tracking Behavioral Changes in Development Programs." ILAC Brief 7 (August), Consultative Group on International Agricultural Research, Montpellier, France. https://cgspace.cgiar.org/bitstream/handle/10568/70174/ILAC_Brief07_mapping.pdf?sequence=1&isAllowed=y.

Wilson-Grau, R. 2018. *Outcome Harvesting: Principles, Steps, and Evaluation Applications*. Charlotte, NC: Information Age Publishing.

Wilson-Grau, R,. and H. Britt. 2012. "Outcome Harvesting." Ford Foundation, Cairo. http://www.managingforimpact.org/sites/default/files/resource/outome_harvesting_brief_final_2012-05-2-1.pdf.

Advanced

Buskens, I., and S. Earl. 2008. "Research for Change: Outcome Mapping's Contribution to Emancipatory Action Research in Africa." *Action Research* 6 (2): 171–92.

Earl, Sarah, and Fred Carden 2002. "Learning from Complexity: The International Development Research Centre's Experience with Outcome Mapping." *Development in Practice* 12 (3–4): 518–24.

Earl, S., F. Carden, and T. Smutylo. 2001. *Outcome Mapping: Building Learning and Reflection into Development Programs*. Ottawa: International Development Research Centre. https://www.idrc.ca/en/book/outcome-mapping-building-learning-and-reflection-development-programs.

Nyangaga, Julius, Terry Smutylo, Dannie Romney, and Patti Kristjanson. 2010. Research that Matters: Outcome Mapping for Linking Knowledge to Poverty-Reduction Actions." *Development in Practice* 20 (8): 972–84.

World Bank. 2014. *Cases in Outcome Harvesting: Ten Pilot Experiences Identify New Learning from Multi-Stakeholder Projects to Improve Results*. Washington, DC: World Bank. https://openknowledge.worldbank.org/handle/10986/20015.

Other Resources

The Outcome Harvesting Community of Practice. https://outcomeharvesting.net/home/.

The Outcome Mapping Learning Community. https://www.outcomemapping.ca/.

The Outcome Mapping Practitioner Guide. https://www.outcomemapping.ca/outcome-mapping-practitioner-guide.

The following articles discuss similar methodologies:

Jabeen, Sumera. 2018. "Unintended Outcomes Evaluation Approach: A Plausible Way to Evaluate Unintended Outcomes of Social Development Programmes." *Evaluation and Program Planning* 68: 262–74.

Lemon, Adrienne, and Mélanie Pinet. 2018. "Measuring Unintended Effects in Peacebuilding: What the Field of International Cooperation Can Learn from Innovative Approaches Shaped by Complex Contexts." *Evaluation and Program Planning* 68: 253–61.

Paz-Ybarnegaray, R., and B. Douthwaite. 2017. "Outcome Evidencing: A Method for Enabling and Evaluating Program Intervention in Complex Systems." American Journal of Evaluation 38 (2): 275–93.

10 Social Network Analysis

BRIEF DESCRIPTION OF THE APPROACH

Social network analysis (SNA) is an approach for examining the relationships (connections) among individuals, groups, organizations, or other entities within a specified network. Its main purpose is to identify and describe the key actors and primary structures that collectively make up a network. In this way, SNA offers a viable method for better understanding the structural links and interrelationships within which programs are embedded and which are created by programs. Compared with system mapping and dynamics, SNA is more focused on the mathematical, structural, or complicated aspects of networks and the number of connections among network elements, rather than their type, quality, or the complex system characteristics that might emerge collectively from microlevel adaptive behavior.

Evaluation questions that SNA may answer include the following:

1. How are individuals, groups, or other entities within a specified network connected?
2. What characterizes the links and the relationships among these individuals and how do they affect the network as a whole?
3. Who might be potential influencers in the network? Who is likely to be influenced?
4. What is the relative importance of a particular institutional actor or individual in a network (for example, a sector, institution, community)?

THE MAIN VARIATIONS OF THE APPROACH

A distinction is often made between two types of network analysis: *ego network analysis and complete network analysis.*

In ego network analysis, each respondent (often as part of a survey) is asked to identify people within their individual network (for example, people they interact with within their village or workplace) and the relationships among these people, providing information about his or her own network. In this type of analysis, no attempt is made to link the individual networks because it is assumed that networks most likely will not overlap (for example, when respondents are sampled randomly). The aim is to assess the characteristics (for example, size, diversity, socioeconomic status) of each respondent's network or to compare the characteristics of the individual with those of her or his network members.

In complete network analysis, all (or almost all) actors (according to particular selection or threshold criteria) within a network are asked to provide information on the people within the specified network they are connected with and their relationships with these individuals. This type of information is cross-checked and verified across respondents and creates a snapshot of the network as a whole: a collective entity that can be visualized (with network maps and graphs) and statistically described. IEOs more often apply SNA to institutions rather than individuals (for example, to analyze the landscape of institutional actors providing support to a particular sector in a country).

THE MAIN PROCEDURAL STEPS OF THE APPROACH

Complete SNA involves five basic steps:

- Defining the network and the individuals who make up the network (this involves defining the boundaries of the network);
- Defining the type of relationship to be examined;
- Collecting relational data from each actor (individual or institutional) in the network, typically by survey;
- Organizing the data in a relational database; and
- Describing and visualizing the network actors and structures. Ego SNA is the same except that the steps are carried out on a subset of network members, so the first step is not needed and the fifth step is not about the network as a whole but about individual (sub)networks.

Data for SNA are often derived from surveys; however, social media and institutional records may also provide useful data. Numerous software packages allow for data management and analysis; some are free of charge, such as Cytoscape, Gephi, and visone.

The data analysis and the description of the actors and their network typically involve different measures of connectivity, including degree centrality (the number of direct connections a respondent holds with other individuals in the network) and closeness centrality (inversely proportional to the average distance between a respondent and all the other respondents in the network). The cohesiveness and density of the network as a whole or of subgroups of the network can also be examined.

THE ADVANTAGES AND DISADVANTAGES OF THE APPROACH

One advantage of SNA is that it offers a systematic approach for documenting and analyzing the interrelationships among the individuals, institutions, or other entities involved in a program, geographical area, or sector. This allows for a better understanding of the structure and functioning of systems, organizational behavior, interorganizational relations, social support, and the flow of information, knowledge, and resources, which helps explain and predict the potential impact of policy changes or implementation processes on relationships among a set of actors. This is particularly relevant for evaluation questions examining supporting or explanatory factors for an outcome. It may serve to focus the evaluation, support a better understanding of the reach of the program, and even provide a context for program theory development.

In some cases, a limitation is the significant investment in time and resources associated with data collection and management. SNA survey questions are often difficult to prepare such that respondents provide reliable answers, and rounds of pilot testing of SNA questions are advised. Application requires analytical skills and the ability to use SNA software. Another limitation is the potential simplification of the relationships depicted in the social networks. By focusing on broad quantifiable patterns, the network maps do not always lend themselves to in-depth analyses of qualitative aspects of the depicted relationships. The addition of qualitative approaches is worth considering. Finally, in complete SNA, the analysis may miss observations. Sometimes it may be difficult to cover the entire relevant population with the data collection instrument (for example, a survey). Coverage must be very high for the SNA to realistically represent, for example, an institutional landscape in a particular area of work.

THE APPLICABILITY OF THE APPROACH

SNAs are becoming more common in development evaluation, and for certain types of evaluation they are feasible and relatively cost efficient. These types of evaluations might deal with topics related to organizational changes, knowledge flows and collaboration, capacity building, and so on. Applications of social network analyses include the following:

1. SNA was applied to understand the positioning of the World Bank Group as a funder and a knowledge leader in the health sector in Liberia.

(*Source*: Vaessen, J., and K. Hamaguchi. 2017. "Understanding the Role of the World Bank Group in a Crowded Institutional Landscape." Blog, Independent Evaluation Group. November 14, 2017. http://ieg.worldbank.org/blog/understanding-world-bank-groups-role.)

2. Network analysis was used to describe how nongovernmental organizations funded by the Ghana Research and Advocacy Project are connected via shared membership in specific issue-related coalitions as part of strategy discussions among these organizations.

   (*Source*: Davies, R. 2009. "The Use of Social Network Analysis Tools in the Evaluation of Social Change Communications." *Monitoring and Evaluation News*. April 2009. https://mande.co.uk/2009/uncategorized/the-use-of-social-network-analysis-tools-in-the-evaluation-of-social-change-communications/.)

3. Network analysis was used to map partner organizations under the Katrine Community Partnerships Program in Uganda.

   (*Source*: Davies, R. 2009. "The Use of Social Network Analysis Tools in the Evaluation of Social Change Communications." *Monitoring and Evaluation News*. April 2009. https://mande.co.uk/2009/uncategorized/the-use-of-social-network-analysis-tools-in-the-evaluation-of-social-change-communications/.)

4. SNA was applied to map policy networks as part of the evaluation of Sexuality Policy Watch, a global forum of organizations and individuals active within the field of sexuality, health, and human rights.

   (*Source*: Drew, R., P. Aggleton, H. Chalmers, and K. Wood. 2011. "Using Social Network Analysis to Evaluate a Complex Policy Network." *Evaluation* 17 (4): 383–94.)

5. SNA was used to examine enterprise networks in regional development projects supported by the European Regional Development Fund.

   (*Source*: Lahdelma, T., and S. Laakso. 2016. "Network Analysis as a Method of Evaluating Enterprise Networks in Regional Development Projects." *Evaluation* 22 (4): 435–50.)

6. Network analysis was used in an evaluation of a multistakeholder water governance initiative in Ghana.

(*Source*: Schiffer, E., and D. Waale. 2008. “Tracing Power and Influence in Networks; Net-Map as a Tool for Research and Strategic Network Planning.” Discussion Paper, International Food Policy Research Institute, Washington, DC. https://www.ifpri.org/publication/tracing-power-and-influence-networks.)

7. Organizational network analysis was used to evaluate and improve the Integration Opportunities for HIV and Family Planning Services in Addis Ababa, Ethiopia.

   (Source: Thomas, J. C., H. Reynolds, C. Bevc, and A. Tsegaye. 2014. “Integration Opportunities for HIV and Family Planning Services in Addis Ababa, Ethiopia: An Organizational Network Analysis.” *BMC Health Services Research* 14, article 22. https://bmchealthservres.biomedcentral.com/track/pdf/10.1186/1472-6963-14-22?site=bmchealthservres.biomedcentral.com.)

8. SNA was used to evaluate whether cluster development programs stimulated the formation of collaborative activities and formal or informal cooperation among cluster members.

   (*Source*: Giuliani, E., and C. Pietrobelli. 2011. “Social Network Analysis Methodologies for the Evaluation of Cluster Development Programs.” Technical Notes IDB-TN-317, Inter-American Development Bank, Washington, DC. https://publications.iadb.org/bitstream/handle/11319/5342/IDB-TN-317%20Social%20Network%20Analysis%20Methodologies%20for%20the%20Evaluation%20of%20Cluster%20Development%20Programs.pdf?sequence=1&isAllowed=y.)

9. SNA was used in Mozambique in an evaluation of the humanitarian assistance after a severe flooding. The aim was to determine how the network structure affected the interorganizational coordination and humanitarian aid outcomes.

   (*Source*: Ramalingam, B. 2006. *Tools for Knowledge and Learning: A Guide for Development and Humanitarian Organizations*. London: Overseas Development Institute. https://www.odi.org/sites/odi.org.uk/files/odi-assets/publications-opinion-files/188.pdf.)

10. SNA was applied to understand trade networks in Gaya, Niger, and its neighboring border cities of Malanville, Benin, and Kamba, Nigeria.

(*Source*: Raj, A., and J-F. Arvis. 2014. “How Social Connections and Business Ties Can Boost Trade: An Application of Social Network Analysis.” Blog, World Bank. April 28, 2014. http://blogs.worldbank.org/trade/how-social-connections-and-business-ties-can-boost-trade-application-social-network-analysis.)

READINGS AND RESOURCES

Background

Borgatti, S. P., M. G. Everett, and J. C. Johnson. 2018. *Analyzing Social Networks*. Thousand Oaks, CA: SAGE.

Kadushin, C. 2012. *Understanding Social Networks: Theories, Concepts, and Findings*. Oxford: Oxford University Press.

Newman, Mark. 2018. *Networks*. Oxford: Oxford University Press.

Scott, John. 2017. *Social Network Analysis*. Thousand Oaks, CA: SAGE.

Wasserman, S., and K. Faust. 1994. *Social Network Analysis: Methods and Applications*. Cambridge, UK: Cambridge University Press.

Advanced

Cadger, K., A. K. Quaicoo, E. Dawoe, and M. E. Isaac. 2016. “Development Interventions and Agriculture Adaptation: A Social Network Analysis of Farmer Knowledge Transfer in Ghana.” *Agriculture* 6 (3): 32.

Cassidy, L., and G. Barnes. 2012. “Understanding Household Connectivity and Resilience in Marginal Rural Communities through Social Network Analysis in the Village of Habu, Botswana.” *Ecology and Society* 17 (4): 11. https://www.jstor.org/stable/26269205.

Crossley, Nick, Elisa Bellotti, Gemma Edwards, Martin G. Everett, Johan Henrik Koskinen, and Mark Tranmer. 2015. *Social Network Analysis for Ego-Nets*. Thousand Oaks, CA: SAGE.

Drew, R., P. Aggleton, H. Chalmers, and K. Wood. 2011. “Using Social Network Analysis to Evaluate a Complex Policy Network.” *Evaluation* 17 (4): 383–94.

Durland, M. M., and K. A. Fredericks, eds. 2005. “Social Network Analysis in Program Evaluation.” Special issue, *New Directions for Evaluation* 2005 (107).

Holman, N. 2008. “Community Participation: Using Social Network Analysis to Improve Developmental Benefits.” *Environment and Planning C: Government and Policy* 26 (3): 525–43.

Johny, Judit, Bruno Wichmann, and Brent M. Swallow. 2017. “Characterizing Social Networks and Their Effects on Income Diversification in Rural Kerala, India.” *World Development* 94: 375–92.

Kolaczyk, E. D., and G. Csárdi. 2014. *Statistical Analysis of Network Data with R*. New York: Springer.

Rudnick, Jessica, Meredith Niles, Mark Lubell, and Laura Cramer. 2019. “A Comparative Analysis of Governance and Leadership in Agricultural Development Policy Networks.” *World Development* 117: 112–26.

Scott, J., and P. Carrington, eds. 2011. *The SAGE Handbook of Social Network Analysis*. Thousand Oaks, CA: SAGE.

Wonodi, C. B., L. Privor-Dumm, M. Aina, A. M. Pate, R. Reis, P. Gadhoke, and O. S. Levine. 2012 “Using Social Network Analysis to Examine the Decision-Making Process on New Vaccine Introduction in Nigeria.” *Health Policy and Planning* 27, suppl. 2 (1): ii27–ii38.

Yang, S., F. B. Keller, and L. Zheng. 2016. *Social Network Analysis: Methods and Examples*. Thousand Oaks, CA: SAGE.

Other Resources

Baker, Matt. 2019. “Demystifying Social Network Analysis in Development: Five Key Design Considerations.” Lab Notes (blog). USAID Learning Lab. March 26, 2019. https://usaidlearninglab.org/lab-notes/demystifying-social-network-analysis-development-five-key-design-considerations.

Davies. R. 2003. “Network Perspectives in the Evaluation of Development Interventions: More Than a Metaphor.” Paper presented at EDAIS Conference, November 24–25, 2003. https://www.mande.co.uk/wp-content/uploads/2003/nape.pdf.

SPECIFIC METHODS: DATA COLLECTION METHODS

This section provides guidance notes on specific methods and tools used in evaluation, often as part of the main methodological approaches presented in the preceding section. We first describe specific methods and tools for *data collection*, followed by specific methods and tools for *data analysis*. This distinction is often blurred in practice, with some methods and tools involving both data collection and analysis. We start the data collection section with literature reviews, which should precede the design of any primary data collection if the latter is to contribute to the expansion of knowledge on a given topic. We then describe qualitative interviews, focus groups, surveys, the Delphi method, scales, and emerging practices and technologies for data collection.

11 Structured Literature Review

BRIEF DESCRIPTION OF THE METHOD

The structured literature review is perhaps best viewed as a survey of existing studies on a given topic. The primary purpose of a structured literature review is to determine the state-of-the-art knowledge on a given topic and provide the best answer to an evaluation or research question, given existing knowledge. The results of a literature review should be the starting point for designing collection of primary data if the latter is to contribute efficiently to the accumulation of knowledge. The defining feature of structured literature reviews is the commitment to systematic and transparent procedures for searching and summarizing existing studies.

Evaluation questions systematic and structured reviews may answer include the following:

1. What do we already know about this topic?
2. How do we discover and synthesize knowledge about this topic?

THE MAIN PROCEDURAL STEPS OF THE METHOD

Structured literature reviews (as often used in IEOs) are relatively "light touch" exercises that follow the main principles of rigorous systematic reviews (as used in academia, but rarely used in IEOs). Systematic reviews typically involve the following four steps:

- Establishing rules to identify the sources that are relevant to the review;
- Establishing stricter rules to determine which of these sources will be included in the review;
- Establishing rules to guide extraction of information from the sources; and
- Establishing rules to synthesize the information extracted.

THE ADVANTAGES AND DISADVANTAGES OF THE METHOD

Structured literature reviews are universally applicable. Different types of structured reviews have different requirements, and the Readings and Resources include sources that represent different traditions of reviews (including the more rigorous and elaborate systematic reviews and realist syntheses). The most obvious advantage of structured literature reviews is that they might answer a question of interest (often relating to effectiveness or impact) or at least provide a platform on which primary data collection can efficiently build. In addition, after clarifying what is already known, we are more confident that data collection can be focused on untapped areas of knowledge and will avoid "reinventing the wheel."

The main shortcoming of structured literature review is perhaps the time needed, depending on the breadth of the existing knowledge on the topic. But this investment is likely to be more efficient than implementing primary data collection without knowing what empirical (evaluation) research already exists.

READINGS AND RESOURCES

Background

Boland, Angela, Gemma Cherry, and Rumona Dickson. 2017. *Doing a Systematic Review: A Student's Guide*. Thousand Oaks, CA: SAGE.

Denyer, D., and D. Tranfield. 2009. "Producing a Systematic Review." In *The SAGE Handbook of Organizational Research Methods*, edited by D. A. Buchanan and A. Bryman, 671–89. Thousand Oaks, CA: SAGE.

Gough, David, Sandy Oliver, and James Thomas. 2017. *An Introduction to Systematic Reviews*. Thousand Oaks, CA: SAGE.

Gough, David, Sandy Oliver, and James Thomas, eds. 2018. *Systematic Reviews and Research. Fundamentals of Applied Research*. Thousand Oaks, CA: SAGE.

Grant, Maria J., and Andrew Booth. 2009. "A Typology of Reviews: An Analysis of 14 Review Types and Associated Methodologies." *Health Information and Libraries Journal* 26 (2): 91–108.

Higgins, Julian P. T., James Thomas, Jacqueline Chandler, Miranda Cumpston, Tianjing Li, Matthew J. Page, and Vivian A. Welch, eds. 2019. *Cochrane Handbook for Systematic Reviews of Interventions*. Oxford: Wiley Blackwell.

Pawson, Ray. 2006. *Evidence-Based Policy: A Realist Perspective*. Thousand Oaks, CA: SAGE.

Pawson, R., T. Greenhalgh, G. Harvey, and K. Walshe. 2005. "Realist Review—A New Method of Systematic Review Designed for Complex Policy Interventions." *Journal of Health Services Research & Policy* 10 (1): 21–34.

Petticrew, Mark, and Helen Roberts. 2005. *Systematic Reviews in the Social Sciences: A Practical Guide*. Oxford: Wiley Blackwell.

Thomas, J., and A. Harden. 2008. Methods for the thematic synthesis of qualitative research in systematic reviews. *BMC Medical Research Methodology* 8: 45.

Advanced

Ananiadou, S., B. Rea, N. Okazaki, R. Procter, and J. Thomas. 2009. "Supporting Systematic Reviews Using Text Mining." *Social Science Computer Review* 27 (4): 509–23.

Anderson, Laurie M., Mark Petticrew, Eva Rehfuess, Rebecca Armstrong, Erin Ueffing, Phillip Baker, Daniel Francis, and Peter Tugwell. 2011. "Using Logic Models to Capture Complexity in Systematic Reviews." *Research Synthesis Methods* 2 (1): 33–42.

Cornish, Flora. 2015. "Evidence Synthesis in International Development: A Critique of Systematic Reviews and a Pragmatist Alternative." *Anthropology & Medicine* 3: 263–77.

Evans, David K., and Anna Popova. 2015. "What Really Works to Improve Learning in Developing Countries? An Analysis of Divergent Findings in Systematic Reviews." Policy Research Working Paper WPS7203, World Bank, Washington, DC.

Higgins, J. P. T., and S. Green, eds. 2011. *Cochrane Handbook for Systematic Reviews of Interventions*, Version 5.1.0. http://handbook.cochrane.org.

Oliver, S., A. Harden, R. Rees, J. Shepherd, G. Brunton, J. Garcia, and A. Oakley. 2005. "An Emerging Framework for Including Different Types of Evidence in Systematic Reviews for Public Policy." *Evaluation* 11 (4): 428–46.

Papaioannou, D., A. Sutton, C. Carroll, A. Booth, and R. Wong. 2010. "Literature Searching for Social Science Systematic Reviews: Consideration of a Range of Search Techniques." *Health Information & Libraries Journal* 27: 114–22.

Pawson, Ray 2001. "Evidence Based Policy: In Search of a Method." Working Paper 3, ESRC UK Centre for Evidence Based Policy and Practice, Swindon, UK. https://www.kcl.ac.uk/sspp/departments/politicaleconomy/research/cep/pubs/papers/assets/wp3.pdf.

Rycroft-Malone, J., B. McCormack, A. M. Hutchinson, Kara DeCorby, Tracey K. Bucknall, Bridie Kent, Alyce Schultz, Erna Snelgrove-Clarke, Cheryl B. Stetler, Marita Titler, Lars Wallin, and Val Wilson.. 2012. "Realist Synthesis: Illustrating the Method for Implementation Research." *Implementation Science* 7: 33.

Snilstveit, B., S. Oliver, and M. Vojtkova. 2012. "Narrative Approaches to Systematic Review and Synthesis of Evidence for International Development Policy and Practice." *Journal of Development Effectiveness* 4 (3): 409–29.

Verhagen, Arianne P., Henrica C. W. de Vet, Robert A. de Bie, Alphons G. H. Kessels, Maarten Boers, Lex M. Bouter, and Paul G. Knipschild. 1998. "The Delphi List: A Criteria List for Quality Assessment of Randomized Clinical Trials for Conducting Systematic Reviews Developed by Delphi Consensus." *Journal of Clinical Epidemiology* 51 (12): 1235–41.

White, Howard, and Hugh Waddington, eds. 2012. "Systematic Reviews." Special issue, *Journal of Development Effectiveness* 4 (3).

Other Resources

Campbell Collaboration. https://campbellcollaboration.org/.

Cochrane Collaboration. https://www.cochranelibrary.com/about/about-cochrane-reviews.

International Initiative for Impact Evaluation. https://www.3ieimpact.org/evidence-hub/publications/systematic-reviews.

Rameses. http://www.ramesesproject.org/Home_Page.php.

Systematic reviews and other review types (Temple University). https://guides.temple.edu/c.php?g=78618&p=3879604.

12 Qualitative Interview

BRIEF DESCRIPTION OF THE METHOD

An interview can be described as a conversation structured around a specific topic or a set of questions—a purposeful conversation. The general purpose of an interview is to better understand a person's experience of a situation, including the person's opinions, feelings, knowledge, information, and perspective. In evaluation, interviews may shed light on program expectations and experiences among staff, stakeholders, or participants.

THE MAIN VARIATIONS OF THE METHOD

There are many variants of interviews. Although most interviews are conducted one on one, group interviews are also common (see guidance note 13, Focus Group). Interviews may be conducted face to face, over the phone, or via online platforms (see guidance note 17, Emerging Technologies for Data Collection).

A common distinction is made between the *unstructured interview*, an open-ended conversation on general topics (used when the purpose is essentially exploratory); the *semistructured interview*, a conversation revolving around a list of topics and questions with varying degrees of specification and a flexible order (the most widely used type of interview); and the *structured interview*, a conversation that stringently adheres to a list of questions with a prespecified order and wording (used to obtain standardized, comparable data across interview participants). Structured interviews are commonly used in surveys (see guidance note 14, Surveys).

THE MAIN PROCEDURAL STEPS OF THE METHOD

The process of designing and conducting interviews usually involves the following six steps:

- Defining the overarching topic for the interviews;
- Identifying possible participants (for example, program participants, program staff, or stakeholders); an interview approach typically uses some type of purposive sampling strategy; for group interviews, consider which subgroups might be relevant (for example, stakeholders from specific institutions, urban versus rural program participants);

- Developing an interview guide (often called a protocol or template) containing the questions or topics to be covered during the interview by considering which type of interview is appropriate (for example, unstructured, semistructured); the guide may also provide guidance to the interviewer about how to conduct the interview;
- Training interviewers on the purpose, content, and procedure of the interviews, and ideally testing or piloting the interview guide to make adjustments;
- Arranging and conducting the sessions (considering whether interviews should be [audio] recorded); and
- Transcribing and analyzing the data.

THE ADVANTAGES AND DISADVANTAGES OF THE METHOD

Individual interviews are particularly relevant for collecting detailed information about how individuals experience a specific situation or, in evaluation, how they experience program activities or even outcomes. The depth and richness of the data may reveal important insights about how and why the program made a difference (or failed to make a difference).

The primary weakness of individual interviewing relates to time and resources. Preparing, arranging, conducting, transcribing, and analyzing interviews is time-consuming. To some extent, the use of computer-assisted software can ease the work burden of interview transcription and analysis (see guidance note 20, Using Computer-Assisted Qualitative Data Analysis Software). It is often difficult to identify and assess the quality of the interviewing process, particularly in unstructured interviews; the interviewer's awareness of biases and factors possibly affecting interviewees' accounts; and the interviewer's ability to adjust the tone and content of the interview while it is in progress. Another commonly cited shortcoming is perhaps best viewed as a necessary trade-off: by focusing on deep and rich information obtained from a select and often small sample of participants, interviews may not lead to conclusive statements about the extent to which the findings are representative of a broader population unless grounded in a carefully constructed purposive sampling approach and subject to triangulation (across interviews, but also with data from other methods).

THE APPLICABILITY OF THE METHOD

Interviews are widely applicable in both development evaluation in general and IEO evaluations specifically. Examples of the use of interviews include the following:

1. Individual interviews and content analysis (a technique for coding and counting specific words and phrases in interview transcripts) were used to learn how villagers perceive the use of hand pumps for water.

   (*Source*: Narayan, D. 1996. “Toward Participatory Research.” Technical Paper 307, World Bank, Washington, DC. http://documents.worldbank.org/curated/en/578241468765924957/pdf/multi0page.pdf.)

2. Interviews were used to identify and understand the mechanisms underlying the implementation of the Accountability for Reasonableness framework in Tanzania.

   (*Source*: Maluka, S., P. Kamuzora, and M. Sansebastian. 2011. “Implementing Accountability for Reasonableness Framework at District Level in Tanzania: A Realist Evaluation.” *Implementation Science* 6: 1–15.)

3. In-depth interviews with key informants, including community leaders, elected representatives, shop owners, energy suppliers, teachers, and health workers were conducted as part of a participatory appraisal of the contribution of energy reforms to poverty reduction in the Republic of Yemen.

   (*Source*: World Bank. 2007. *Tools for Institutional, Political, and Social Analysis of Policy Reform: A Sourcebook for Development Practitioners*. Washington, DC: World Bank, 229–32. http://documents.worldbank.org/curated/en/434581468314073589/Tools-for-institutional-political-and-social-analysis-of-policy-reform.)

4. Open-ended interviews with key informants were used to inform the development of a baseline survey in an evaluation of a tea sector reform in Rwanda.

   (*Source*: World Bank. 2007. *Tools for Institutional, Political, and Social Analysis of Policy Reform: A Sourcebook for Development Practitioners*. Washington, DC: World Bank, 227–9. http://documents.worldbank.org/curated/en/434581468314073589/Tools-for-institutional-political-and-social-analysis-of-policy-reform.)

READINGS AND RESOURCES

Background

Brinkmann, Svend, and Steinar Kvale. 2014. *Interviews: Learning the Craft of Qualitative Research Interviewing.* Thousand Oaks, CA: SAGE.

Brinkmann, Svend, and Steinar Kvale. 2018. *Doing Interviews*. Thousand Oaks, CA: SAGE.

Edwards, R. 2013. *What Is Qualitative Interviewing?* London and New York: Bloomsbury. http://eprints.ncrm.ac.uk/3276/1/complete_proofs.pdf; https://www.cdc.gov/healthyyouth/evaluation/pdf/brief17.pdf.

King, Nigel, Christine Horrocks, and Joanna Brooks. 2018. *Interviews in Qualitative Research*. Thousand Oaks, CA: SAGE.

Manzano, A. 2016. "The Craft of Interviewing in Realist Evaluation." *Evaluation* 22 (3): 342–60.

Patton, M. Q. 2015. *Qualitative Research and Evaluation Methods: Integrating Theory and Practice*, 4th ed. Thousand Oaks, CA: SAGE.

Advanced

Boyce, Carolyn, and Palena Neale 2006. "Conducting In-Depth Interviews: A Guide for Designing and Conducting In-Depth Interviews for Evaluation Input." Pathfinder International Tool Series: Monitoring and Evaluation—2, Pathfinder International, Watertown, MA. http://www2.pathfinder.org/site/DocServer/m_e_tool_series_indepth_interviews.pdf.

Kallio, H., A. Pietila, M. Johnson, and M. Kangasniemi. 2016. "Systematic Methodological Review: Developing a Framework for a Qualitative Semi-Structured Interview Guide." *Journal Of Advanced Nursing* 72 (12): 2954–65. http://usir.salford.ac.uk/id/eprint/39197.

Mammen, J. R., S. A. Norton, H. Rhee, and A. M. Butz. 2016. "New Approaches to Qualitative Interviewing: Development of a Card Sort Technique to Understand Subjective Patterns of Symptoms and Responses." *International Journal of Nursing Studies* 58: 90–96.

Oltmann, Shannon. 2016. "Qualitative Interviews: A Methodological Discussion of the Interviewer and Respondent Contexts." *Forum Qualitative Sozialforschung / Forum: Qualitative Social Research* 17 (2-S.1), article 15. http://www.qualitative-research.net/index.php/fqs/article/view/2551/3998.

Seidman, Irving. 2019. *Interviewing as Qualitative Research: A Guide for Researchers in Education and the Social Sciences*, 5th ed. New York: Teachers College Press.

Sewell, M. n.d. "The Use of Qualitative Interviews in Evaluation." CYFERnet—Evaluation, University of Arizona. https://cals.arizona.edu/sfcs/cyfernet/cyfar/Intervu5.htm.

13 Focus Group

BRIEF DESCRIPTION OF THE METHOD

The focus group is a type of group interview where the evaluator facilitates a structured discussion to one or more specific topics—a purposeful group conversation. This allows for multiple respondents (usually five to ten individuals) to formulate, discuss, and refine their individual and collective perspectives and experiences. The purpose of focus groups is primarily to collect information about similarities and differences in experiences and perspectives among specific groups of people.

In evaluation, focus groups are commonly used to collect information on similarities and differences in experiences with program implementation, outcomes, or both among program staff, participants, or stakeholders. However, focus groups are widely applicable and can also be used for needs assessment or program theory development, for example.

THE MAIN VARIATIONS OF THE METHOD

The practical implementation of focus groups may vary significantly depending on the specific purpose and context. Focus groups may be characterized by different degrees of structure (from highly structured to free flowing), the use of different prompts to facilitate conversation (for example, photos, program materials, or findings from previous data collection), and the extent to which the facilitator engages in the discussion (from passive observation to active participation).

THE MAIN PROCEDURAL STEPS OF THE METHOD

The design and implementation of focus groups usually involves six steps:

Defining the overarching topic for the focus groups;

Identifying relevant groups of participants (for example, program participants); focus groups typically use some type of purposive sampling strategy; consider which subgroups should be included (for example, stakeholders from specific types of institutions, urban versus rural program participants);

- Developing a topic guide listing the most salient topics to be covered during the focus group, and perhaps instructions to the facilitator on conducting the focus group;
- Training the facilitators on the purpose, content, and procedures of the focus group;
- Recruiting and arranging the participants in relevant subgroups for each focus group and conducting these sessions (taking into account local cultural conditions, which may include separation of participants according to ethnicity, gender, or culturally specific status differentials); and
- Transcribing and analyzing the data.

THE ADVANTAGES AND DISADVANTAGES OF THE METHOD

A significant strength of focus groups is that they may allow for diverse views to be shared and discussed, whereby new insights and opinions may emerge and take form. This quality information may generate insights and findings difficult to obtain through other, less interactive, data collection methods. As such, focus groups also lend themselves well to collaborative evaluation approaches (see guidance note 7, Participatory Approaches to Evaluation). Moreover, members of some vulnerable groups may feel more comfortable in expressing their views in the presence of their peers rather than in one-on-one interviews with an external evaluator.

Focus groups may have several significant shortcomings. A group may be influenced by uneven group dynamics (for example, an assertive participant dominating the discussion), which in turn may influence the findings. A group discussion may center on socially desirable perspectives and experiences, as participants may be more comfortable describing and emphasizing positive rather than negative program experiences. If participants are selected by a local organization, there is a risk people with a certain political affiliation may be included without the knowledge of the evaluator. Some respondent groups may be reluctant or simply unable to participate (for example, participants with higher levels of responsibility in an institution). The practical feasibility of the criteria for selecting respondents and the adherence to these is important to consider and ensure. Successfully moderating focus groups typically involves greater preparation and skill than is generally required for conducting one-on-one interviews. Finally, like interviews, focus groups may be subject to the shortcomings of selectivity and generalizability of evidence.

THE APPLICABILITY OF THE METHOD

Focus groups have been applied in a broad range of development evaluations and are widely used by IEOs. Examples include the following:

1. The evaluation of five urban development projects in Brazil involved focus groups with beneficiaries (grouped according to responses to a prior survey) from project and nonproject cities.

   (*Source*: White, H. 2006. “Impact Evaluation: The Experience of the Independent Evaluation Group of the World Bank.” Working Paper 38268, World Bank, Washington, DC. http://documents.worldbank.org/curated/en/475491468138595632/Impact-evaluation-the-experience-of-the-independent-evaluation-group-of-the-World-Bank.)

2. Focus groups were used as part of a World Bank study on the use of public transportation in Lima, Peru. By allocating male and female participants to separate focus groups, important gender differences on priorities and barriers for use of public transportation were identified.

   (*Source*: Bamberger, M., J. Rugh, and L. Mabry. 2006. *RealWorld Evaluation: Working under Budget, Time, Data, and Political Constraints, 2nd ed.* Thousand Oaks, CA: SAGE.)

3. A team designing an HIV/AIDS activity in Kenya conducted focus group discussions with potential target groups and service providers to gain a deeper understanding of various issues and constraints related to the epidemic.

   (*Source*: UK Department for International Development. 2003. *Tools for Development: A Handbook for Those Engaged in Development Activity*, version 15.1. London: DFID. http://webarchive.nationalarchives.gov.uk/+/http:/www.dfid.gov.uk/Documents/publications/toolsfordevelopment.pdf.)

4. Focus groups were used as part of an ex post assessment of the impact of closing selected Agricultural Development and Marketing Corporation markets in Malawi.

   (*Source*: World Bank. 2007. *Tools for Institutional, Political, and Social Analysis of Policy Reform. A Sourcebook for Development Practitioners*. Washington, DC: World Bank. https://siteresources.worldbank.org/EXTTOPPSISOU/Resources/1424002-1185304794278/TIPs_Sourcebook_English.pdf.)

READINGS AND RESOURCES

Background

Barbour, Rosaline. 2018. *Doing Focus Groups*. Thousand Oaks, CA: SAGE.

Cyr, Jennifer. 2019. Focus Groups for the Social Science Researcher. Cambridge, UK, and New York: Cambridge University Press.

Guichard, Anne, Émilie Tardieu, Christian Dagenais, Kareen Nour, Ginette Lafontaine, and Valéry Ridde. 2017. "Use of Concurrent Mixed Methods Combining Concept Mapping and Focus Groups to Adapt a Health Equity Tool in Canada." *Evaluation and Program Planning* 61: 169–77.

Krueger, R. A. 2002. "Designing and Conducting Focus Group Interviews." Eastern Illinois University. https://www.eiu.edu/ihec/Krueger-FocusGroupInterviews.pdf.

Krueger. Richard. 2014. *Focus Groups: A Practical Guide for Applied Research*. Thousand Oaks, CA: SAGE.

Morgan, D. L. 1988. *Focus Group as Qualitative Research*. Thousand Oaks, CA: SAGE.

Ryan, K. E., T. Gandha, M. J. Culberson, and C. Carlson. 2014. "Focus Group Evidence: Implications for Design and Analysis." *American Journal of Evaluation* 35 (3): 328–45.

Sewell, M. n.d. "Using Focus Groups for Evaluation." CYFERnet—Evaluation, University of Arizona. https://cals.arizona.edu/sfcs/cyfernet/cyfar/focus.htm.

Stewart, David W., and Prem N. Shamdasani. 2014. *Focus Groups: Theory and Practice*. Thousand Oaks, CA: SAGE.

Advanced

Allen, M. D. 2014. "Telephone Focus Groups: Strengths, Challenges, and Strategies for Success." *Qualitative Social Work* 13 (4): 571–83.

Barbour, Rosaline S., and David L. Morgan. 2017. *A New Era in Focus Group Research: Challenges, Innovation and Practice*. London: Palgrave Macmillan.

Chen, J., and P. Neo. 2019. "Texting the Waters: An Assessment of Focus Groups Conducted via the WhatsApp Smartphone Messaging Application." *Methodological Innovations* 12 (3).

Farnsworth, John, and Bronwyn Boon. 2010. "Analyzing Group Dynamics within the Focus Group." *Qualitative Research* 10 (5): 605–24.

Morgan, David L. 2019. *Basic and Advanced Focus Groups*. Thousand Oaks, CA: SAGE.

Onwuegbuzie, A. J., W. B. Dickinson, N. L. Leech, and A. G. Zoran. 2009. "A Qualitative Framework for Collecting and Analyzing Data in Focus Group Research." *International Journal of Qualitative Methods* 8 (3): 1–21.

Nyumba, Tobias O., Kerrie Wilson, Christina J. Derrick, and Nibedita Mukherjee 2018. “The Use of Focus Group Discussion Methodology: Insights from Two Decades of Application in Conservation.” *Methods in Ecology and Evolution* 9 (1): 20–32.

Winship, Gary, and Julie Repper. 2007. “Focus Group Research: The Role of Cacophony in Dialogical Democracy.” *Group Analysis* 40 (1): 125–39.

14 Survey

BRIEF DESCRIPTION OF THE METHOD

Surveys are one of the most common tools for data collection. The primary purpose of surveys is to collect information (often, but not always, quantitative) on a specified sample of respondents in a systematic and consistent manner. For evaluations, surveys can be useful for determining the distribution of characteristics or outcomes in a sample or population of interest and documenting differences and similarities in characteristics and outcomes across subgroups of a sample or population, or even over time.

THE MAIN VARIATIONS OF THE METHOD

Across the many variations in survey practice, a distinction can be made between *cross-sectional surveys* (where information is collected from a sample at one point in time) and *longitudinal surveys* (where information is collected from a sample at different points in time). The latter, especially if collecting information on the same sample, allows for analysis of changes over time.

Surveys may involve two broad types of questions: *open-ended questions*, where the respondent is asked to provide any response to a specific question prompt (for example, "How did you experience the program?"); and *closed-ended questions*, where the respondent is asked to respond by selecting one or more predefined response categories (for example, "satisfied" or "unsatisfied"). Although open-ended questions serve well to yield deeper, more nuanced, and potentially unexpected answers and insights, they are more time-consuming for both the respondent and the evaluator. In contrast, closed-ended questions with short single-answer responses lessen the burden for data processing and analysis significantly.

Surveys may be administered face to face, through phone interviews, by mail, by email or a designated website via computer (or smartphones), and via short message service (SMS) with (regular) mobile phones.

THE MAIN PROCEDURAL STEPS OF THE METHOD

Designing a survey usually involves the following seven steps:

- Clarifying the survey's purpose and identifying the overarching issues and questions to be answered by the survey;
- Defining the (reference) population and corresponding sample of respondents;
- Developing the questionnaire (see guidance note 16, Developing and Using Scales);
- Conducting a pilot test and refining the questionnaire accordingly;
- Administering the survey to the specified sample;
- Entering the data (if not done automatically as in electronic and tablet-based surveys) and cleaning the data; and
- Analyzing and reporting the findings.

THE ADVANTAGES AND DISADVANTAGES OF THE METHOD

There are many advantages and disadvantages to surveys. Some of the most common are mentioned here.

One methodological advantage of the survey method is that respondents may answer the same set of predetermined questions, allowing for collection of comparable data across large numbers of people (and over time). Moreover, the use of random sampling supports a broad range of statistical analyses and (under certain assumptions) allows for statistical inference of findings. In this way, surveys are useful for collecting generalizable data over a relatively short time.

Some key limitations of surveys relate to the type of data they gather and the possibility of bias. First, surveys are best suited for collecting quantitative data (that is, quantifiable responses). Although this type of information supports a broad range of statistical analyses, it may fail to capture deeper and more fine-grained nuances of the respondents' experiences and perspectives. The use of open-ended questions may remedy this concern, but the data they yield are generally more time-consuming and difficult to analyze. Second, responses to survey questions are easily influenced by the specific wording, sequence, and type of questions; the structure and wording of the response categories provided; and other design features of the questionnaire, potentially biasing the findings. Questionnaire development, although seemingly simple, is a difficult task and central to the quality of any survey. Nonresponse, and especially, nonrandom nonresponse may significantly complicate the generalizability of findings (and the use of statistical inference).

THE APPLICABILITY OF THE METHOD

Surveys are widely applicable and relevant for collecting information on trends and patterns across a large group of individuals. Surveys are widely used in development evaluation in general and IEO evaluations specifically. Examples include the following:

1. Repeated surveys of a stratified sample of households were used as part of the evaluation of the feeder road program (under the Eritrean community development fund) to capture the road's effect on nearby households.

   (*Source*: Bamberger, M., J. Rugh, and L. Mabry. 2020. 3rd ed. *RealWorld Evaluation: Working under Budget, Time, Data, and Political Constraints*. Thousand Oaks, CA: SAGE.)

2. Repeated surveys were used among a stratified random sample of households in Bangalore, India, to collect data on service use related to telephones, electricity, water and sewerage, public hospitals, transportation, and banks.

   (*Source*: World Bank. 2004. *Influential Evaluations: Evaluations That Improved Performance and Impacts of Development Programs.* International Bank for Reconstruction and Development. Washington, DC: World Bank. https://www.ircwash.org/sites/default/files/Bamberger-2004-Influential.pdf.)

3. Household surveys were used to obtain initial data on the availability and use of ration shops as part of an evaluation of wheat flour ration shops in Pakistan.

   (*Source*: World Bank. 2004. *Influential Evaluations: Evaluations That Improved Performance and Impacts of Development Programs*. International Bank for Reconstruction and Development. Washington, DC: World Bank. https://www.ircwash.org/sites/default/files/Bamberger-2004-Influential.pdf.)

4. The 2015/16 Ethiopia Rural Socioeconomic Survey combined household and community surveys to document agricultural and livestock practices, harvest outcomes, and socioeconomic conditions.

(*Source*: World Bank. 2017. “Ethiopia—Socioeconomic Survey 2015–2016 Wave 3.” Data set. World Bank, Washington, DC.https://datacatalog.worldbank.org/dataset/ethiopia-socioeconomic-survey-2015-2016 and https://microdata.worldbank.org/index.php/catalog/2783.)

5. The successive Living Standards Measurement Study covers many topics—for example, Integrated Surveys on Agriculture in Burkina Faso, Niger, Ethiopia, Nigeria, Malawi, Tanzania, Mali, and Uganda covered agriculture, socioeconomic status, and nonfarm income activities.

   (*Source*: World Bank. 2015. “LSMS—Integrated Surveys on Agriculture.” World Bank, Washington, DC. http://surveys.worldbank.org/lsms/programs/integrated-surveys-agriculture-ISA and http://siteresources.worldbank.org/INTLSMS/Resources/3358986-1233781970982/5800988-1282216357396/7337519-1388758418241/GHS_Panel_Survey_Report_Wave_2.pdf.)

READINGS AND RESOURCES

Background

Blair, Johnny Jr., Ronald F. Czaja, and Edward Blair. 2013. *Designing Surveys: A Guide to Decisions and Procedures*. Thousand Oaks, CA: SAGE.

de Leeuw, E. D., J. J. Hox, and D. A. Dillman. 2008. *The International Handbook of Survey Methodology*. Jena: European Association of Methodology. http://joophox.net/papers/SurveyHandbookCRC.pdf.

Fink, Arlene. 2015. *How to Conduct Surveys: A Step-by-Step Guide*. Thousand Oaks, CA: SAGE.

Fowler, Floyd 2013. *Survey Research Methods*. Thousand Oaks, CA: SAGE.

Glewwe, P. 2005. *Household Sample Surveys in Developing and Transition Countries*. Studies in Methods. Department of Economic and Social Affairs, Statistics Division. New York: United Nations. https://unstats.un.org/unsd/hhsurveys/pdf/household_surveys.pdf.

Grosh, M., and P. Glewwe. 2000. *Designing Household Survey Questionnaires for Developing Countries: Lessons from 15 Years of the Living Standards Measurement Study*. Washington, DC: World Bank Group. http://documents.worldbank.org/curated/en/452741468778781879/Volume-One.

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

World Bank. https://openknowledge.worldbank.org/bitstream/handle/10986/6975/350340The0Powe1n0REV01OFFICIAL0USE1.pdf.

Rea, Louis, and Richard Parker 2014. *Designing and Conducting Survey Research: A Comprehensive Guide*. San Francisco: Jossey-Bass.

Robinson, Sheila B., and Kimberly Firth Leonard. 2019. *Designing Quality Survey Questions.* Thousand Oaks, CA: SAGE.

Advanced

Dillman, Don A., Jolene D. Smyth, and Leah Melani Christian. 2014. *Internet, Phone, Mail, and Mixed-Mode Surveys: The Tailored Design Method*. Hoboken, NJ: Wiley.

Harkness, Janet A., Michael Braun, Brad Edwards, Timothy P. Johnson, Lars E. Lyberg, Peter Ph. Mohler, Beth-Ellen Pennell, and Tom W. Smith, eds. 2010. *Survey Methods in Multinational, Multiregional, and Multicultural Contexts*. Hoboken, NJ: Wiley.

Kalton, Graham 2020. *Introduction to Survey Sampling*. Thousand Oaks, CA: SAGE.

Litwin, M. 2003. *How to Assess and Interpret Survey Psychometrics*, 2nd ed. The Survey Kit. Thousand Oaks, CA: SAGE.

Ruel, Erin, III, William E. Wagner, and Brian Joseph Gillespie. 2015. *The Practice of Survey Research: Theory and Applications*. Thousand Oaks, CA: SAGE.

Johnson, Timothy P., Beth-Ellen Pennell, Ineke A. L. Stoop, and Brita Dorer, eds. 2018. *Advances in Comparative Survey Methods: Multinational, Multiregional, and Multicultural Contexts* (3MC). Wiley Series in Survey Methodology. Hoboken, NJ: Wiley.

Other Resources

Demographic and Health Surveys are nationally representative household surveys that provide data for a wide range of indicators in population, health, and nutrition. https://dhsprogram.com/What-We-Do/Survey-Types/DHS.cfm.

Doing Business provides objective measures of business regulations and their enforcement across 190 economies and selected cities at the subnational and regional level. https://www.doingbusiness.org/en/methodology.

Freedom in the World is an annual study of political rights and civil liberties around the world. https://freedomhouse.org/report/freedom-world.

The Global Findex database, the world’s most comprehensive data set on how adults save, borrow, make payments, and manage risk, has been updated every three years since 2011 and includes nationally representative surveys of more than 150,000 adults in over 140 economies. https://globalfindex.worldbank.org/.

The Living Standards Measurement Study is a household survey program housed within the Survey Unit of the World Bank’s Development Data Group that maintains a repository of questionnaires and interviewer manuals. http://surveys.worldbank.org/.

15 Delphi Method

BRIEF DESCRIPTION OF THE METHOD

The Delphi method is a systematic and iterative process for eliciting opinions and determining consensus among a broad range of stakeholders. The method can be used to solicit feedback and reach consensus on a broad range of topics, issues, or needs. Findings from the Delphi method are particularly relevant for planning, problem solving, decision-making, program development, and forecasting.

THE MAIN VARIATIONS OF THE METHOD

Practical applications of the Delphi method vary in the use of paper-and-pencil versus online administration of the questionnaires, the number of iterative rounds used to solicit opinions, and the extent to which these rounds are structured or unstructured, among other things. A more recent variant, the face-to-face Delphi method, uses small group conferences.

THE MAIN PROCEDURAL STEPS OF THE METHOD

The Delphi procedure commonly involves the following five steps:

- Selecting the stakeholders to be included;
- Designing and administering a questionnaire eliciting stakeholder opinions on one or more specified topics;
- Analyzing the findings and summarizing an initial group response to the topics;
- Developing and administering a refined questionnaire eliciting stakeholder opinions on the group responses; and
- Analyzing and summarizing a refined group response. The fourth and fifth steps can be repeated until a final group consensus is reached.

THE ADVANTAGES AND DISADVANTAGES OF THE METHOD

One advantage of the Delphi method is that it allows for consensus building among geographically dispersed stakeholders without the organizational constraints and expenses of an in-person meeting. The survey format allows individual stakeholders to share their opinions anonymously and with equal status while allowing them to reflect on and respond directly to the group consensus. In this way, the method allows for group interaction without direct confrontation between stakeholders with opposing opinions. In some cases, a Delphi panel can constitute a reasonable alternative to a structured literature review in areas where very little literature exists (for example, on perspectives regarding the future of a particular sector).

There are also disadvantages to the method. The iterative process demands sustained commitment among respondents. If reaching consensus involves multiple rounds of data collection, the process can be lengthy. The final consensus is highly influenced by the framing of the questions and statements in the questionnaire, the number of rounds, and the stakeholders included. Accordingly, these aspects and their potential implications should be carefully considered in the design of the Delphi method. The quality of evidence generated from the Delphi panel depends greatly on the levels of participants' knowledge and the selection of participants (influencing potential generalizability).

THE APPLICABILITY OF THE METHOD

The Delphi method is applicable to a wide range of issues and topics and easily adapted in its administration for specific stakeholder needs. Practical applications of the Delphi method include the following:

1. A Delphi panel was used to elicit global experts' opinions on trends in the renewable energy sector in the framework of the Independent Evaluation Group's evaluation on Bank Group support for renewable energy.

   (*Source*: Jayawardena, Migara, Enno Heijndermans, Maurya West Meiers, Ryan Watkins, Joy Butscher, Shenghui Feng, and Noureddine Berrah. Forthcoming. "An Oracle for the Future of Renewable Energy: Global Experts Predict Emerging Opportunities and Challenges through a Delphi Technique." Independent Evaluation Group, World Bank, Washington, DC).

2. The Delphi method was used to solicit expert opinion and identify consensus on agrifood policy options.

(*Source*: Frewer, L. J., A. R. H. Fischer, M. T. A. Wentholt, H. J. P. Marvin, B. W. Ooms, D. Coles, and G. Rowe. 2011. "The Use of Delphi Methodology in Agrifood Policy Development: Some Lessons Learned." *Technological Forecasting & Social Change* 78: 1514–25.

3. The Delphi method was used as part of an environmental valuation of the Amazon rain forest, where environmental valuation experts from several countries were asked to predict willingness to pay for different preservation options.

   (*Source*: Strand, J., R. T. Carson, S. Navrud, A. Ortiz-Bobea, and J. Vincent. 2014. "A 'Delphi Exercise' as a Tool in Amazon Rainforest Valuation." Policy Research Working Paper 7143, World Bank, Washington, DC. https://openknowledge.worldbank.org/bitstream/handle/10986/21139/WPS7143.pdf?sequence=1.)

4. A Delphi survey was used as part of USAID's Ending Preventable Child and Maternal Deaths strategy to identify priority health areas and develop consensus on a prioritized research agenda.

   (*Source*: USAID [US Agency for International Development]. n.d. "USAID Social and Behavior Change Programs for Ending Preventable Child and Maternal Deaths." USAID, Washington, DC. http://pdf.usaid.gov/pdf_docs/PBAAH599.pdf.)

5. A Delphi survey among 60 international malaria research experts was used to identify research priorities and elicit relative valuations of the potential impact of different types of health research.

   (*Source*: Mulligan, J.-A., and L. Conteh. 2016. "Global Priorities for Research and the Relative Importance of Different Research Outcomes: An International Delphi Survey of Malaria Research Experts." *Malaria Journal* 15: 585. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5139033/.)

6. A Delphi method was used to develop a competency-based tool for evaluation of community-based training in undergraduate medical education in India.

   (*Source*: Shewade, H. D., K. Jeyashree, S. Kalaiselvi, C. Palanivel, and K. C. Panigrahi. 2017. "Competency-Based Tool for Evaluation of Community-

Based Training in Undergraduate Medical Education in India—A Delphi Approach." *Advances in Medical Education and Practice* 8: 277–86. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5395246/.)

READINGS AND RESOURCES

Background

Day, Jacqueline, and Bobeva, Milena. 2005. "A Generic Toolkit for the Successful Management of Delphi Studies." *The Electronic Journal of Business Research Methods* 3 (2): 103–16.

Hasson, Felicity, Sinead Keeney, and Hugh McKenna. 2000. "Research Guidelines for the Delphi Survey Technique." *Journal of Advanced Nursing* 32 (4): 1008–15.

Hsu, C.-C., and B. A. Sandford. 2007. "The Delphi Technique: Making Sense of Consensus." *Practical Assessment, Research, and Evaluation* 12, article 10. https://scholarworks.umass.edu/pare/vol12/iss1/10/.

Linstone, H. A., and M. Turoff, eds. 2002. *The Delphi Method—Techniques and Applications*. Newark: New Jersey Institute of Technology. https://web.njit.edu/~turoff/pubs/delphibook/delphibook.pdf.

Rowe, Gene, and George Wright. 2011. "The Delphi Technique: Past, Present, and Future Prospects—Introduction to the Special Issue." *Technological Forecasting and Social Change* 78 (9): 1487–90.

Advanced

Christie, Christina A., and Eric Barela. 2005. "The Delphi Technique as a Method for Increasing Inclusion in the Evaluation Process." *The Canadian Journal of Program Evaluation* 20 (1): 105–22.

Erffmeyer, R. C., E. S. Erffmeyer, and I. M. Lane. 1986. "The Delphi Technique: An Empirical Evaluation of the Optimal Number of Rounds." *Group & Organization Studies* 11 (1–2): 120–8.

Garavalia, L., and M. Gredler. 2004. "Teaching Evaluation through Modeling: Using the Delphi Technique to Assess Problems in Academic Programs." *American Journal of Evaluation* 25 (3): 375–80.

Geist, Monica R. 2010. "Using the Delphi Method to Engage Stakeholders: A Comparison of Two Studies." *Evaluation and Program Planning* 33 (2): 147–54.

Giannarou, Lefkothea, and Efthimios Zervas. 2014. "Using Delphi Technique to Build Consensus in Practice." *International Journal of Business Science and Applied Management* 9 (2).

Hsu, Chia-Chien, and Brian A Sandford. 2007. “Minimizing Non-Response in the Delphi Process: How to Respond to Non-Response.” *Practical Assessment, Research, and Evaluation* 12, article 17.

Hung, H.-L., J. W. Altschuld, and Y. F. Lee. 2008. “Methodological and Conceptual Issues Confronting a Cross-Country Delphi Study of Educational Program Evaluation.” *Evaluation and Program Planning* 31 (2): 191–8.

Jiang, Ruth, Robin Kleer, and Frank T. Piller. 2017. “Predicting the Future of Additive Manufacturing: A Delphi Study on Economic and Societal Implications of 3D Printing for 2030.” *Technological Forecasting and Social Change* 117: 84–97.

Yousuf, Muhammad Imran. 2007. “Using Experts’ Opinions through Delphi Technique.” *Practical Assessment, Research, and Evaluation* 12, article 4. https://scholarworks.umass.edu/pare/vol12/iss1/4.

16 Developing and Using Scales

BRIEF DESCRIPTION OF THE METHOD

A scale in its simplest form is a set of survey questions that collectively measure a specific concept (for example, poverty, self-esteem, farmer resilience). In evaluation, scaling is particularly relevant when data on outcomes are collected in the form of attitudes (satisfaction), emotions (quality of life), or even motivational states (self-efficacy), to name but a few. These are concepts (sometimes referred to as constructs) that are not directly observable and often broad in meaning, allowing for many different interpretations. Accordingly, scale development can be a useful step when developing surveys capturing these types of concepts.

THE MAIN VARIATIONS OF THE METHOD

As an alternative to scale development, the use of existing scales should be considered if relevant and appropriate—a broad and growing range of scales is available. One practical benefit of using an existing scale is avoiding the often time-consuming task of defining and operationalizing a given concept. However, in the selection of a relevant scale, its appropriateness should be carefully considered in terms of its relevance (does it match the informational needs of the evaluation?) and feasibility (can it practically be administered as part of the evaluation?). Moreover, the scale should be relevant and appropriate for the intended respondents in the evaluation.

THE MAIN PROCEDURAL STEPS OF THE METHOD

Developing a scale requires the following six steps:

- Defining the concept to be measured (specifying the different subdimensions that collectively make up the concept);
- Formulating a corresponding set of questions and answer categories that match the definition of the concept;
- Conducting a pilot test of the scale (Are the questions difficult to understand? Do the response options match the question?);
- Analyzing the validity and reliability of the scale (using statistical analysis);

- Modifying the scale based on the pilot test results; and
- Administering the developed scale to a broader sample of respondents or observations.

Whether a new or existing scale is used, the scale chosen should always be characterized by careful correspondence between each survey question and a specific aspect of the concept being measured; use of clear and simple language (avoid complex terms and slang); avoidance of items that all respondents would uniformly endorse (either positively or negatively); avoidance of questions that include more than one topic (double-barreled questions).

THE ADVANTAGES AND DISADVANTAGES OF THE METHOD

Scales are relevant in most evaluations. For example, project performance may be assessed using ordinal scales, each scale representing a level of performance. If applied appropriately, the use of a validated scale may enhance the validity (precision) and reliability (consistency) of the evaluation findings (see appendix A, Glossary of Key Terms, for definitions of these terms). Moreover, the use of existing scales may allow for comparisons or at least guidance on scoring or rating interpretations. There are also drawbacks. Developing a scale can be time-consuming, and pilot testing especially demands a considerable amount of time and effort. The validation of the scale demands some level of statistical expertise. Finally, the precision of the scale should match the degree of differentiation in evidence that can meaningfully support a particular scoring or rating.

THE APPLICABILITY OF THE METHOD

The use of scales is highly applicable in development evaluation and IEO evaluations specifically. Examples include the following:

1. The Resilience Index Measurement and Analysis Model II measures the physical and capacity dimensions of household resiliency to food security shocks.

   (*Source*: FAO [Food and Agriculture Organization of the United Nations]. 2016. *RIMA-II Resilience Index Measurement and Analysis*. Rome: FAO. https://reliefweb.int/sites/reliefweb.int/files/resources/AnalysIng%20Resilience%20for%20better%20targeting%20and%20action.pdf.)

2. The Human Development Index, the Inequality-Adjusted Human Development Index, the Gender Development Index, the Gender Inequality Index, and the Multidimensional Poverty Index are composite measures of average achievement in key dimensions of human development. These are developed and administered by the United Nations Development Programme, and data and summary statistics on the measures are available at the country level.

   (*Source*: United Nations Development Programme. 2017. "Global Human Development Indicators." http://hdr.undp.org/en/countries.)

3. The Integrated Questionnaire for the Measurement of Social Capital—with a focus on applications in developing countries—aims to provide quantitative data on various dimensions of social capital, including groups and networks, trust and solidarity, collective action and cooperation, information and communication, social cohesion and inclusion, and empowerment and political action. Developed by the World Bank, the tool has been pilot tested in Albania and Nigeria and has been widely used.

   (*Source*: Grootaert, C., D. Narayan, M. Woolcock, and V. Nyhan-Jones. 2004. "Measuring Social Capital: An Integrated Questionnaire." Working Paper 18, World Bank, Washington, DC. http://documents.worldbank.org/curated/en/515261468740392133/Measuring-social-capital-an-integrated-questionnaire.)

4. The Women's Empowerment in Agriculture Index, launched in February 2012 by the International Food Policy Research Institute, Oxford Poverty and Human Development Initiative, and USAID's Feed the Future, is a comprehensive and standardized measure that captures women's empowerment and inclusion in the agricultural sector.

   (*Source*: International Food Policy Research Institute. 2012. "WEAI Resource Center." IFPRI, Washington, DC. http://www.ifpri.org/topic/weai-resource-center.)

READINGS AND RESOURCES

Background

Arora, L., and W. Trochim. 2013. "Rating Systems in International Evaluation." i-eval Think Piece 3, Evaluation Unit, International Labour Organization, Geneva.

http://www.ilo.org/wcmsp5/groups/public/---ed_mas/---eval/documents/publication/wcms_202430.pdf.

Bandalos, Deborah L. 2017. *Measurement Theory and Applications for the Social Sciences*. New York: Guilford Press.

DeVellis, R. F. 2017. *Scale Development: Theory and Applications. Applied Social Research Methods*. Thousand Oaks, CA: SAGE.

Johnson, Robert L., and Grant B. Morgan. 2016. *Survey Scales: A Guide to Development, Analysis, and Reporting*. New York: Guilford Press.

Advanced

Alsop, Ruth, and Nina Heinsohn. 2005. "Measuring Empowerment in Practice: Structuring Analysis and Framing Indicators." Policy Research Working Paper 3510, World Bank, Washington, DC. https://openknowledge.worldbank.org/handle/10986/8856.

Fernald, Lia C. H., Elizabeth Prado, Patricia Kariger, and Abbie Raikes. 2017. *A Toolkit for Measuring Early Childhood Development in Low and Middle-Income Countries*. Strategic Impact Evaluation Fund. Washington, DC: World Bank. https://openknowledge.worldbank.org/handle/10986/29000.

Greco, Salvatore, Alessio Ishizaka, Menelaos Tasiou, and Gianpiero Torrisi. 2019. "On the Methodological Framework of Composite Indices: A Review of the Issues of Weighting, Aggregation, and Robustness." *Social Indicators Research* 141 (1): 61–94.

Litwin, M. 2003. How to Assess and Interpret Survey Psychometrics, 2nd ed. *The Survey Kit*. Thousand Oaks, CA: SAGE.

Morgado, F. F. R., J. F. F. Meireles, C. M. Neves, A. C. S. Amaral, and M. E. C. Ferreira. 2018. "Scale Development: Ten Main Limitations and Recommendations to Improve Future Research Practices." *Psicologia: Reflexão e Crítica* 30, article 3.

Rowan, Noell, and Dan Wulff. 2007. "Using Qualitative Methods to Inform Scale Development." *Qualitative Report* 12 (3): 450–66. https://files.eric.ed.gov/fulltext/EJ800203.pdf.

Sturgess, P. 2016. "Measuring Resilience." Evidence on Demand, UK Department for International Development, London. https://www.gov.uk/dfid-research-outputs/measuring-resilience.

17 Emerging Technologies for Data Collection

BRIEF DESCRIPTION OF THE METHOD

The rapid growth in the availability and accessibility of information and communication technology has opened up innovative avenues for data collection relevant to evaluation. Enlisting these technologies for data collection offers three advantages: (i) generating efficiencies in existing modes of operationalization; (ii) providing access to new sources of data; and (iii) expanding the reach of relevant information to remote and developing areas.

THE MAIN VARIATIONS OF THE METHOD

A comprehensive description of the vast and growing array of emerging technologies for data collection is beyond the scope of this guide. However, several applications are worth highlighting. *Mobile devices* such as smartphones and tablets have emerged as a widely accessible means of facilitating real-time data collection through SMS surveys, data collection apps, or "crowdsourced" reporting (where participants upload responses to an open-source software platform). Mobile devices have also been used in financial transactions, reducing the time and cost to issue consumer receipts and increasing tax compliance. In countries where more people have mobile devices than computers, mobile devices have proven useful for generating early-warning disaster notifications. Such devices have several clear advantages for data collection in evaluation. First, they are less expensive than computers, and are therefore more commonly available for data collection. Second, they offer greater mobility, granting access to more remote areas. Third, various built-in functions such as location sensing allow practitioners greater access to a broader spectrum of metadata for decision-making and resource management, including location monitoring.

Another—closely related—avenue for data collection and management is the use of web-based software platforms to store real-time data collected from multiple users. To illustrate, software such as Sensemaker allows for real-time collection and analysis of respondent narratives. Crowdsourced data (for example, survey responses, audio, images) can also be combined with GPS-based geographic information to facilitate the analysis of geographical trends and variations.

The increased use of mobile devices has also resulted in *digital traces* (*big data*) that can be collected and analyzed for evaluative purposes. These include postings and activities on social media (for example, Twitter and Facebook) and metadata involving location, movement, and social interaction. With the advent of new forms of data analytics these digital traces can be sifted for meaning, documenting behavioral trends in social media, the progression of epidemics, access and use of infrastructural assets, and transportation patterns, among other things (see guidance note 19, Emerging Applications in Data Science and Data Analytics).

Another emerging technology is the use of *drones* (also referred to as unmanned aerial vehicles) for data collection, mapping, and monitoring in remote, hard-to-reach areas. Equipped with cameras, drones can collect aerial photographs and videos; create maps; chart topographical features, land cover, roads, buildings, and other infrastructure; and detect changes over time. *Geospatial data* have also been combined with geocoded data on development programs. This type of aid mapping can be used to survey the geographical dispersion of development investments, coordinate development aid efforts, generate data for improving own-source revenue generation, expand property registration and tax compliance, enable more responsive monitoring and asset management systems, and provide context for impact evaluations of development projects.

Finally, *satellite imagery*, which is publicly available through the NASA Earth Observations website, for example, has been used to estimate economic development and activity and regional growth, and, in combination with machine learning, to predict average household wealth and expenditures, educational attainment, or health outcomes in rural areas (see guidance note 19, Emerging Applications in Data Science and Data Analytics).

THE MAIN PROCEDURAL STEPS OF THE METHOD

Though the breadth of this approach renders generation of a single set of procedural steps difficult, five general guidelines can be mapped:

- Determining the type of data to be measured (focusing on the unit of analysis and the specific data collection needs);
- Formulating a data collection strategy to take advantage of the efficiencies provided by one or more of the emerging forms of data collection explored;
- Accounting for the potential weaknesses of the data collection plan, including a strategy for the treatment of missing values (specifically nonrandom gaps in the data resulting from various idiosyncrasies of the data collection plan used);

- Selecting a consistent and unbiased method of analysis for the collected data; and
- Assessing the potential for scaling and replicability of the method of data collection being considered.

THE ADVANTAGES AND DISADVANTAGES OF THE METHOD

Using emerging technologies can (i) increase the feasibility and reach of the data collection; (ii) minimize delays in data entry and processing; (iii) lower the cost of data collection; (iv) expand access to remote areas; and (v) provide real-time or at least more readily updated data for analysis. Moreover, texts, images, or even audio and video recordings can be collected through mobile devices, potentially broadening the types of data available for the evaluation. A greater diversity of data sources can improve the quality (for example, the credibility) of the evaluation, allowing practitioners to triangulate among multiple sources to assess the impact of complex programs.

Although emerging technologies can expand the reach and efficiency of available data, evaluators must ensure that overreliance on available information does not lead to "convenience sampling," whereby conclusions are based not on the reference population but rather on the (biased) sample of available data. These technologies can offer evaluators hitherto untapped troves of data, but data collection also must be designed to ensure that the samples used are both representative of and meaningful for the subject of analysis.

A related concern is that overreliance on remote data collection may limit the evaluators' understanding of the context within which data were collected. Moreover, owing to the evolving and unstructured nature of the information collected, it may be difficult (if not impossible) to assess its quality. Exclusive reliance on mobile devices may introduce selection bias by excluding those without devices or the skills to use them or those in areas with limited internet or cellular connectivity.

Finally, there are a number of important ethical issues and regulatory frameworks to consider when using emerging technologies. These include protocols for acquiring respondents' consent, privacy rights, proper anonymization of metadata and aggregate population demographics, ethical codes of conduct, customs and aviation regulations (for drone use), and telecom or data protection regulations, among others. Accordingly, the use of emerging technologies should always involve a careful assessment and mitigation of potential consequences for affected communities, and comply with relevant regulations and permissions. Given disparities in national guidelines and regulations related to mobile data collection, digital traces, geographic information systems, and online sampling, both ethical and regulatory considerations should be given full and measured consideration in the design phase of any data collection approach.

THE APPLICABILITY OF THE METHOD

Applications of emerging technologies for data collection—although growing—are still relatively rare in both development evaluation and IEO evaluations. Examples include the following:

1. The BRAC organization (formerly Bangladesh Rural Advancement Committee) contacted almost 12,000 village-level organizations in Bangladesh to ask community members what their priorities were. Its frontline staff workers took advantage of regular meetings in communities to conduct a poll and send in community priorities by SMS (May, 2013).

   (*Source*: Raftree, L., and M. Bamberger. 2014. *Emerging Opportunities: Monitoring and Evaluation in a Tech-Enabled World*. New York: The Rockefeller Foundation. https://www.rockefellerfoundation.org/wp-content/uploads/Monitoring-and-Evaluation-in-a-Tech-Enabled-World.pdf.)

2. The United Nations Children's Fund worked with local partners in Uganda to engage over 100,000 young people as u-Reporters who then received SMS polls.

   (*Source*: UNICEF. 2012. "U-Report Application Revolutionizes Social Mobilization, Empowering Ugandan Youth." https://www.unicef.org/french/infobycountry/uganda_62001.html.)

3. The Oxfam America Horn of Africa Risk Transfer for Adaptation weather-indexed crop insurance program in Ethiopia used a program theory and collected satellite-generated rainfall data as part of its monitoring and evaluation system.

   (*Source*: Raftree, L., and M. Bamberger. 2014. *Emerging Opportunities: Monitoring and Evaluation in a Tech-Enabled World*. New York: The Rockefeller Foundation. https://www.rockefellerfoundation.org/wp-content/uploads/Monitoring-and-Evaluation-in-a-Tech-Enabled-World.pdf.)

4. The PartoPen is a multimedia pen and tablet used for maternal health training, data recording, and built-in monitoring. The PartoPen system provides training instructions, task reminders, and audio feedback in real time. The system also can detect abnormal labor progression by analyzing data and provides audio and text-based feedback to encourage birth attendants to take appropriate action. Evaluators used the tool's capabilities to measure errors, corrections, and information recorded and to evaluate whether providing a tutorial on the PartoPen improved health workers' use of it.

   (*Source*: Underwood, H., S. Sterling, and J. K. Bennett. 2013. "The PartoPen in Training and Clinical Use—Two Preliminary Studies in Kenya." *Proceedings of the International Conference on Health Informatics* 1: 112–21.)

5. A web-based platform, Sensemaker, was used in Rwanda to collect parents' perceptions of a mentoring program. The platform allowed simultaneous data entry (including narratives) by multiple respondents in different locations.

   (*Source*: Raftree, L., and M. Bamberger. 2014. *Emerging Opportunities: Monitoring and Evaluation in a Tech-Enabled World*. New York: The Rockefeller Foundation. https://www.rockefellerfoundation.org/wp-content/uploads/Monitoring-and-Evaluation-in-a-Tech-Enabled-World.pdf.)

6. As part of the Secondary Education Improvement Project in Ghana, a smartphone platform combining crowdsourcing of citizen input (images and testimonies) with GPS-based geographic information was used for mapping and real-time monitoring of schools.

   (*Source*: World Bank. 2017. "Ghana Secondary Education Improvement Project." World Bank, Washington, DC. http://projects.worldbank.org/P145741?lang=en.)

7. Satellite imagery of densities of buildings and cars, the prevalence of shadows, road lengths, roof materials, agricultural crops, and other textural features was used to predict several poverty indicators of the Household Income Expenditure Survey and the Census of Population and Housing.

   (*Source*: Donaldson, D., and A. Storeygard. 2016. "The View from Above: Applications of Satellite Data in Economics." *Journal of Economic Perspectives* 30 (4): 171–98.)

8. Aid mapping, using a publicly available data set of 61,243 World Bank project locations, was used to estimate regional impact variations on vegetative carbon sequestration.

   (*Source*: Runfola, D., A. B. Yishay, J. Tanner, G. Buchanan, J. Nagol, M. Leu, S. Goodman, R. Trichler, and R. Marty. 2017. “A Top-Down Approach to Estimating Spatially Heterogeneous Impacts of Development Aid on Vegetative Carbon Sequestration.” *Sustainability* 9 (3): 409.)

9. Impact evaluation using mobile devices has been particularly effective in Sub-Saharan Africa. For example, the Competitive African Cotton Initiative, the African Cashew Initiative, and the Coffee Project Tanzania all took advantage of mobile devices (specifically tablets) to conduct impact evaluations and help increase local farm revenues in Ghana, Côte d’Ivoire, and Tanzania.

   (*Source*: Leidig, Mathias, Richard M. Teeuw, and Andrew D. Gibson. 2016. “Data Poverty: A Global Evaluation for 2009 to 2013—Implications for Sustainable Development and Disaster Risk Reduction.” *International Journal of Applied Earth Observation and Geoinformation* 50: 1–9. https://www.sciencedirect.com/science/article/abs/pii/S0303243416300241.)

10. The Kenya Municipal Program relied on data collected via smartphones and tablets to generate demographic and economic estimates for 15 cities in Kenya, drawing on data from over 150,000 households and more than 5,000 variables to assess the quality of and access to key infrastructure.

    (*Source*: World Bank. 2010. “Kenya Municipal Program.” World Bank, Washington, DC. https://projects.worldbank.org/en/projects-operations/project-detail/P066488?lang=en.)

11. The Geo-Enabling Initiative for Monitoring and Supervision takes advantage of an open-source data tool kit (based on the Harvard Humanitarian Initiative’s free KoBo Toolbox) to generate a centralized monitoring and evaluation system, helping build capacity in areas where connectivity and access are limited because of conflict, fragility, or violence.

    (*Source*: World Bank. 2019. *Geo-Enabling Initiative for Monitoring and Supervision*. Washington, DC: World Bank Group. http://documents.worldbank.org/curated/en/271431561153010274/Geo-Enabling-Initiative-for-Monitoring-and-Supervision-GEMS.)

READINGS AND RESOURCES

Background

Bamberger, M. 2016. *Integrating Big Data into the Monitoring and Evaluation of Development Programmes*. New York: UN Global Pulse. http://unglobalpulse.org/sites/default/files/IntegratingBigData_intoMEDP_web_UNGP.pdf.

Raftree, L., and M. Bamberger. 2014. *Emerging Opportunities: Monitoring and Evaluation in a Tech-Enabled World*. New York: The Rockefeller Foundation. https://www.rockefellerfoundation.org/wp-content/uploads/Monitoring-and-Evaluation-in-a-Tech-Enabled-World.pdf.

Jäckle, A., A. Gaia, and M. Benzeval. 2018. *The Use of New Technologies to Measure Socioeconomic and Environmental Concepts in Longitudinal Studies*. London: CLOSER. https://www.closer.ac.uk/wp-content/uploads/CLOSER-resource-The-use-of-new-technology-to-measure-socio-economic-and-environmental-concepts.pdf.

Lee, Jae-Gil, and Minseo Kang. 2015. "Geospatial Big Data: Challenges and Opportunities." *Big Data Research* 2 (2): 74–81.

Li, Songnian, Suzana Dragicevic, Francesc Antón Castro, Monika Sester, Stephan Winter, Arzu Coltekin, Christopher Pettit, Bin Jiang, James Haworth, Alfred Stein, and Tao Cheng. 2016. "Geospatial Big Data Handling Theory and Methods: A Review and Research Challenges." *ISPRS Journal of Photogrammetry and Remote Sensing* 115: 119–33.

UNDP. 2013. "Innovations in Monitoring & Evaluating Results." Discussion Paper, UNDP, New York. http://www.undp.org/content/undp/en/home/librarypage/capacity-building/discussion-paper--innovations-in-monitoring---evaluating-results.html.

Advanced

BenYishay, Ariel, Daniel Runfola, Rachel Trichler, Carrie Dolan, Seth Goodman, Bradley Parks, Jeffery Tanner, Silke Heuser, Geeta Batra, and Anupam Anand. 2017. "A Primer on Geospatial Impact Evaluation Methods, Tools, and Applications." AidData Working Paper 44, AidData at William & Mary, Williamsburg, VA. https://www.aiddata.org/publications/a-primer-on-geospatial-impact-evaluation-methods-tools-and-applications.

Bruce, K., and A. Koler. 2016. "Applying Emergent Technologies to Complex Program Evaluation from the INGO Perspective." In *Dealing with Complexity in Development Evaluation*, edited by M. Bamberger, J. Vaessen, and E. Raimondo. Thousand Oaks, CA: SAGE.

Finucane, Mariel McKenzie, Ignacio Martinez, and Scott Cody. 2018. "What Works for Whom? A Bayesian Approach to Channeling Big Data Streams for Public Program Evaluation." *American Journal of Evaluation* 39 (1): 109–22.

Hilton, Lara G., and Tarek Azzam. 2019. "Crowdsourcing Qualitative Thematic Analysis." *American Journal of Evaluation* 40 (4).

Houston, J. Brian, Joshua Hawthorne, Mildred F. Perreault, Eun Hae Park, Marlo Goldstein Hode, Michael R. Halliwell, Sarah E. Turner McGowen, Rachel Davis, Shivani Vaid, Jonathan A. McElderry, and Stanford A. Griffith. 2015. "Social Media and Disasters: A Functional Framework for Social Media Use in Disaster Planning, Response, and Research." *Disasters* 39 (1): 1–22.

Jäckle, Annette, Jonathan Burton, Mick P. Couper, and Carli Lessof. 2019. "Participation in a Mobile App Survey to Collect Expenditure Data as Part of a Large-Scale Probability Household Panel: Coverage and Participation Rates and Biases." *Survey Research Methods* 13 (1): 23–44.

Jacobson, Miriam R., Cristina E. Whyte, and Tarek Azzam. 2018. "Using Crowdsourcing to Code Open-Ended Responses: A Mixed Methods Approach." *American Journal of Evaluation* 39 (3): 413–29.

Joshi, Anirudha, Mandar Rane, Debjani Roy, Nagraj Emmadi, Padma Srinivasan, N. Kumarasamy, Sanjay Pujari, Davidson Solomon, Rashmi Rodrigues, D. G. Saple, Kamalika Sen, Els Veldeman, and Romain Rutten. 2014. "Supporting Treatment of People Living with HIV/AIDS in Resource Limited Settings with IVRs." *CHI '14: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1595–1604.

Heipke, Christian. 2010. "Crowdsourcing Geospatial Data." *ISPRS Journal of Photogrammetry and Remote Sensing* 65 (6): 550–7.

Leidig, Mathias, Richard M. Teeuw, and Andrew D. Gibson. 2016. "Data Poverty: A Global Evaluation for 2009 to 2013—Implications for Sustainable Development and Disaster Risk Reduction." *International Journal of Applied Earth Observation and Geoinformation* 50: 1–9.

Other Resources

Frontier Technologies Hub works with UK Aid to apply frontier technologies to the biggest challenges in development. https://medium.com/frontier-technology-livestreaming/.

ICTworks is the premier community for international development professionals committed to using new and emerging technologies to help communities accelerate their social and economic development. https://www.ictworks.org/.

Insight2Impact provides a collection of blog posts on innovative data collection methods and issues. https://i2ifacility.org/search?tag_term=data-collection.

Principles for Digital Development is a website created and maintained by a community of practice of professionals working in digital development. https://digitalprinciples.org.

Spatial Agent: A World of Data at Your Fingertips. https://olc.worldbank.org/about-olc/spatial-agent-world-data-your-fingertips.

United Nations—Big Data for Sustainable Development Portal. https://www.un.org/en/sections/issues-depth/big-data-sustainable-development/index.html.

The United Nations Integrated Geospatial Information Framework provides a basis and guide for developing, integrating, strengthening, and maximizing geospatial information management and related resources in all countries. https://ggim.un.org/IGIF/.

UK Department for International Development Digital Strategy 2018 to 2020: Doing Development in a Digital World. https://www.gov.uk/government/publications/dfid-digital-strategy-2018-to-2020-doing-development-in-a-digital-world.

SPECIFIC METHODS: DATA ANALYSIS METHODS

This section provides guidance notes on specific methods and tools for data analysis. These are commonly used in combination with the main methodological approaches and specific methods and tools for data collection presented in the preceding sections of this chapter. Most data analysis methods require either qualitative or quantitative data and can be used in combination with several different evaluation approaches: for example, case studies can draw on the analysis of both qualitative and quantitative data, and quasi-experiments rely on theoretical assumptions that might need literature reviews or qualitative data. However, some approaches require their own specific data formats or strategies for data analysis: for example, QCA is applied to Boolean or fuzzy data sets, and process tracing assesses the probative value of data strictly in the context of specific assumptions; this value might change completely under different assumptions and contexts. This section will cover only data analysis methods that are relatively neutral in this sense and can be used with a variety of approaches.

We start by recapitulating the typical, “traditional” ways of analyzing quantitative data sets and then discuss how emerging tools in data science and data analytics are expanding the realm of what is possible in this field. We conclude by covering computer-based tools used to analyze data typically collected during interviews and focus group discussions.

Notice that qualitative data analysis as a broad concept is not limited to the analysis of interview transcripts: this guide does not cover observation and desk reviews, but these are integral to data collection and analysis in evaluation and research. Some approaches mentioned in this guide, such as process tracing, also make heavy use of timelines and particular kinds of observations (sometimes known as trace evidence) that are considered conclusive because they uniquely arise under the assumption that the theory under investigation is true.

18 Conducting Statistical Analysis with “Traditional” Data Sets

BRIEF DESCRIPTION OF THE METHOD

Secondary analysis is the analysis of data or information that was gathered by someone else (for example, researchers, government agencies), gathered for some other purpose, or a combination of these. Data collected by others are termed *secondary data* or *archival data*; *primary data* are collected directly by the evaluation team. Data sets providing useful and accessible information are increasingly available (see Other Resources in guidance note 14, Surveys).

THE MAIN VARIATIONS OF THE METHOD

The volume of data generated and openly shared by research organizations and institutes, public administrations and agencies, and nonprofit organizations has increased immeasurably. These data include demographic and economic data from household surveys and censuses, employment and wage statistics from government databases, and health and poverty data (for example, using standardized indices and indicators), to name but a few (see guidance note 14, Surveys).

Depending on the informational needs and the nature of the data, a broad range of statistical analyses can be conducted with secondary sources of data, including point estimates for selected populations (for example, to establish a baseline; see guidance notes 2, Experimental Approaches, and 3, Quasi-Experimental Approaches), cross-tabulations and proportions (for example, to determine distributions), analysis of variance and other significance tests across subpopulations (for example, to compare specific characteristics or outcomes), multivariate statistics (for example, multiple regression models), and time series analyses (for example, capturing trends over time). These analyses are mostly conducted for descriptive or causal reasons. For description, for example, they can help in understanding how given resources are distributed within a population of interest and what characteristics of individuals, households, firms, or geographical areas they correlate with (see guidance note 14, Surveys). For causation, they aim to answer causal questions and often inform the reconstruction of counterfactual or comparison situations (see guidance note 3, Quasi-Experimental Approaches).

THE MAIN PROCEDURAL STEPS OF THE METHOD

Using secondary data sources should, at a bare minimum, involve the following four steps:

- Defining the type of data needed for the evaluation;
- Exploring existing and reliable sources for relevant secondary data;
- Examining the background, quality, and properties of the identified data; and
- Extracting and analyzing the data using the appropriate (statistical) methods.

THE ADVANTAGES AND DISADVANTAGES OF THE METHOD

The use of secondary data can be both valuable and cost-effective. Statistical analysis of secondary data can be used for better understanding of context and patterns (for example, trend analysis). This is particularly the case for aggregate indicators that represent high-level signals of relevant issues and a barometer of changing conditions that an evaluation could take into account. In retrospective evaluations, secondary data can also be used for establishing a baseline and inform the identification of comparison groups for counterfactual-based causal inference. Further discussion on the use of (secondary) data to determine the net effect of a program can be found in guidance notes 2, Experimental Approaches, and 3, Quasi-Experimental Approaches.

The use of secondary data also comes with limitations. One all-too-common issue is that secondary data, especially large-scale databases assembled for administrative purposes, may not match the informational needs of specific evaluations. The scope and level of aggregation may leave the data difficult to anchor in a local context. Other limitations typically involve nontransparent and potentially unreliable data collection procedures, and time gaps in dissemination and accessibility.

THE APPLICABILITY OF THE METHOD

The increasing availability of secondary data sets makes the approach increasingly accessible and applicable in both development evaluation in general and IEO evaluations more specifically. Practical applications of secondary (statistical) data analysis include the following:

1. An analysis was conducted as part of the Independent Evaluation Group's Shared Prosperity evaluation, where the association between World Bank development policy lending and the quality of the client governments' social policies was examined, using the Country Policy and Institutional Assessment ratings.

   (*Source*: World Bank. 2017. *Growth for the Bottom 40 Percent: The World Bank Group's Support for Shared Prosperity*. Independent Evaluation Group. Washington, DC: World Bank. https://ieg.worldbankgroup.org/evaluations/shared-prosperity.)

2. Existing data sets were used to gather evidence related to the contribution of the Bank Group to furthering regional integration over the evaluation period, estimating a macro-level difference-in-differences model of regional integration in the trade and transport sectors.

   (*Source*: World Bank. 2019. *Two to Tango: An Evaluation of World Bank Group Support to Fostering Regional Integration*. Independent Evaluation Group. Washington, DC: World Bank. https://ieg.worldbankgroup.org/evaluations/regional-integration.)

3. An analysis examined relationships among Bank Group interventions and reduced costs of trade, along with a set of control variables identified in the relevant literature as important codeterminants (for example, quality of infrastructure, quality of institutions, and stage of economic development).

   (*Source*: World Bank. 2019. *Grow with the Flow: An Independent Evaluation of World Bank Group Support to Facilitating Trade 2006–17*. Independent Evaluation Group. Washington, DC: World Bank. https://ieg.worldbankgroup.org/evaluations/facilitating-trade.)

READINGS AND RESOURCES

(See also guidance notes 2, Experimental Approaches, 3, Quasi-Experimental Approaches, and 14, Surveys.)

Background

Bruce, Peter, and Andrew Bruce. 2017. *Practical Statistics for Data Scientists: 50 Essential Concepts*. Newton, MA: O'Reilly Media.

Field, Andy. 2020. *Discovering Statistics Using IBM SPSS*. Thousand Oaks, CA: SAGE.

Frankfort-Nachmias, Chava, and Anna Leon-Guerrero. 2018. *Social Statistics for a Diverse Society*, 8th ed. Thousand Oaks, CA: SAGE.

Advanced

Field, Andy, Jeremy Miles, and Zoë Field. 2012. *Discovering Statistics Using R*. Thousand Oaks, CA: SAGE.

Heeringa, S. G., B. T. West, and P. A. Berglund. 2017. *Applied Survey Data Analysis*. New York: Chapman and Hall/CRC.

Wickham, Hadley, and Garrett Grolemund. 2016. *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data*. Newton, MA: O'Reilly Media.

19 Emerging Applications in Data Science and Data Analytics

BRIEF DESCRIPTION OF THE METHOD

Data science is an interdisciplinary field focused on extracting knowledge from data sets, which are typically large. The field encompasses analysis, preparing data for analysis, and presenting findings to inform high-level decisions in an organization. As such, it incorporates skills from computer science, mathematics, statistics, information visualization, graphic design, and business. In 2015, the American Statistical Association identified database management, statistics and machine learning, and distributed and parallel systems as the three emerging foundational professional communities. There is no broad consensus on the definition of data science, but, at least historically, it is a development of traditional statistics. This is a rapidly evolving field; spurred by fast-growing technological capacity to access and process big data, applications in data science and analytics revolve around harvesting and using the digital traces generated by social media, mobile devices, and online search engines, among other sources. In addition, it includes text analytics (text as data) and machine learning. Explorative, descriptive, and predictive analyses based on these digital traces are but a few examples of how data science and data analytics can complement evaluation.

THE MAIN VARIATIONS OF THE METHOD

The vast and ever-expanding array of variations in data science and data analytics makes a comprehensive overview of evaluation applications difficult, if not impossible. Broadly speaking, it usually involves the integration of large volumes of data, sometimes from multiple sources (some of which are collected in real time); the use of statistical modeling and machine learning to identify patterns (associations) in the data; and the presentation of these using data summarization and visualization techniques. In development evaluation, these tools can be applied to a broad variety of data, including, for example, evaluation reports (that is, text as data), electronic transactions, social media data, satellite imagery, and phone records. For example, using satellite imagery, mobile phone use, and financial transaction data as proxies for poverty measurement can be very efficient but requires “ground-truthing” with other types of data (for example, survey data) to assess consistency.

Due to their prominence in emerging evaluation practice, three broad approaches for using text analytics and machine learning in evaluation will be explored in this guidance note: holistic summarization, classification and extraction, and content analysis. These represent general categories of analysis, with different degrees of depth and variations in applications.

Several tools can be harnessed for *summarization* and holistic portfolio evaluation. Various models of text analytics (including, among others, correlated topic modeling, latent Dirichlet allocation, cluster analysis of term-correlation associations, n-gram analysis, sentiment analysis, and network models) can be leveraged to generate efficiencies. Such approaches begin with a relatively modest goal, seeking to summarize common concepts, constructs, and variations in a large corpus of work, drawing on machine learning tools that streamline descriptive analysis in a robust and systematic way. A combination of descriptive tools can give evaluators a broad sense of the common themes and associations inherent in a large body of text. They are particularly useful as a first step in the evaluation of thematic work, helping to focus evaluators' priors on the commonalities shared across texts. Used alongside conventional qualitative reviews and expert appraisals, these tools can assist evaluators in identifying the universe of cases relevant to an evaluation (that is, identifying the evaluand), supplementing theory-driven insights, and capturing minute variations that might otherwise be missed in a manual review of, for example, project documents.

Relatedly, *classification and extraction* involve emerging methods that can help mine useful information and sort it according to either user-inputted (supervised) or machine-generated (unsupervised) categories. Classification can be useful for both diagnostic and reference purposes: unlike manual approaches, these tools allow evaluators to capture the full variation in program descriptions or outcomes. This provides a distinct advantage over a qualitative-only approach, which can suffer from incompleteness, selection effects due to incomplete or biased sampling, and anecdotal insights. For unsupervised learning and classification, methods such as K-means clustering, Word2Vec modeling, latent Dirichlet allocation, and cluster dendograms can be used to identify broad categories within a corpus of text. Pictorial n-grams can generate a visual guide to common word frequencies, aiding in additional automated categorization or manual classification. These tools are most effective when compared with human-coded labels to ensure high convergence between the two. The use of supervised learning methods can ensure greater accuracy in classification, though they require more time and effort than unsupervised methods. These methods typically "train" a model based on a subset of hand-coded tags, allowing evaluators to input useful theoretical priors and practical experience into the classification procedure. For instance, multicategory classification can be achieved via a multinomial Bayes classifier, though its effectiveness will be a direct function of the quality and size of the training set on which it is based.

Finally, there are innovations relevant for *content analysis*. Tools such as sequence-to-sequence models offer a supervised option to generate summaries from relatively uniform input data based on finite vocabulary and relatively common word-phrase combinations. However, such applications tend to be less useful in evaluation, owing simply to the innate complexity of the subjects under consideration. There exists a real trade-off between efficiency and utility in this respect: although automated methods for content analysis can parse and summarize data, there is a risk that much of the underlying nuance will be lost, even with a relatively large training set. These risks are further compounded when unsupervised abstractive tools are used without proper cross-validation and triangulation with manual analysis. Factual inaccuracy and self-repetition can be endemic, although work and progress in this field continues. One broad variant of text analytics is social media analytics. This includes analyses of publicly available content from social media and professional networking sites (for example, Facebook, Twitter, LinkedIn). To illustrate, LinkedIn Talent Insights, a self-service, big data analytics program, allows for both local and global analyses of job and employment patterns and educational attainment profiles, and can even map skill growth and distribution in specific geographical locations.

THE MAIN PROCEDURAL STEPS OF THE METHOD

Three preliminary steps can be defined in the applications of text analytics and machine learning to evaluation:

- Identifying a well-defined research question applicable to this type of analysis;
- Conducting a review of the inventory of available and applicable (textual) data; and
- Preparing the data for analysis, accounting for patterns of systematically missing values, and reflecting on other biases in the data. After these steps, the specific needs of the evaluation must be considered. If the application requires purely descriptive methods, the data could be treated with a combination of unsupervised machine learning methods and basic statistical summarization or visualization techniques. Where extractive (such as categorization) or content analysis is required, a combination of unsupervised and supervised learning methods (with subsequent cross-validation and statistical analysis) is appropriate. This requires an iterative approach, relying on a sufficiently informative training data set and reliability tests comparing outputs from automated content analysis to manual coding.

THE ADVANTAGES AND DISADVANTAGES OF THE METHOD

The application of data science and data analytics in evaluation is promising in many ways. Machine learning and data analytics offer three significant advantages relative to manual approaches: (i) streamlining the analysis of large volumes of data; (ii) capturing and updating information in real time; and (iii) conducting a holistic review of underlying features (see guidance note 17, Emerging Technologies for Data Collection). Data science and data analytics can complement evaluation in a number of different ways. In evaluation, these applications can be used for various purposes: *descriptive* (for example, providing information on the nature of the program and of the context within which a program operates—before, during, or after program implementation), *extractive* (for example, categorizing text based on common features), or even *prescriptive* (for example, using simulation models and forecasting techniques to advise on how to redesign the program for specific outcomes). When combined with appropriate statistical modeling techniques, such applications can lend greater analytical rigor to evaluations, assessing correlates of program effectiveness, diagnosing common failures, and assessing complex associations for better understanding of underlying trends and patterns in the data.

Although emerging analytical tools for text analytics can venture into more sophisticated approaches such as sentiment analysis or topic modeling, they should be seen not as a substitute for but as a complement to traditional (desk review) practices in evaluation. Absent proper supervision and practitioner inputs, such models are liable to generate biased, inconsistent, or even meaningless information. To illustrate, patterns in internet traffic may reflect the behavior of actual users, computer programs (also known as bots), or some combination of these. Moreover, the internet users generating the data may not be representative of other individuals who have limited or no access to the internet, potentially introducing selectivity bias. Finding the right balance between modeling and cross-validation is therefore important, a task that is as much a science as an art. Best practices require multiple iterations of data analysis paired with successive user input to guarantee the overall validity of the output generated. Close collaboration between evaluators and data scientists is key to guaranteeing the quality of the analysis and its findings.

THE APPLICABILITY OF THE METHOD (BEYOND TEXT ANALYTICS)

The increased accessibility and capacity to use big data makes data science and data analytics increasingly applicable in both development evaluation in general and IEO evaluations specifically. Examples include the following:

1. The United Nations Population Fund, as part of a population census in Afghanistan, developed population maps using geographic information system modeling to integrate demographic survey data with satellite imagery and other infrastructure data.

   (*Source*: Bamberger, M. 2016. *Integrating Big Data into the Monitoring and Evaluation of Development Programmes*. New York: UN Global Pulse. http://unglobalpulse.org/sites/default/files/IntegratingBigData_intoMEDP_web_UNGP.pdf.)

2. Translator Gator, a language game, was developed to create text mining dictionaries for recognizing sustainable development–related conversations in Indonesia across the local dialects, jargon, and alphabets. In 2017, Global Pulse released Translator Gator 2 to test whether crowdsourcing can be used to develop taxonomies for disaster-related keywords to inform disaster management efforts.

   (*Source*: Bamberger, M. 2016. *Integrating Big Data into the Monitoring and Evaluation of Development Programmes*. New York: UN Global Pulse. http://unglobalpulse.org/sites/default/files/IntegratingBigData_intoMEDP_web_UNGP.pdf.)

3. CycloMon is an analytics and visualization platform developed to assist governments in helping communities prepare for and respond to the impact of tropical cyclones. The platform was developed to monitor social response before, during, and after cyclones across 14 countries in the Pacific region.

   (*Source*: Bamberger, M. 2016. *Integrating Big Data into the Monitoring and Evaluation of Development Programmes*. New York: UN Global Pulse. http://unglobalpulse.org/sites/default/files/IntegratingBigData_intoMEDP_web_UNGP.pdf.)

4. Financial transaction data can be used to better understand the economic resilience of people affected by natural disasters. An example is a project that measured daily point-of-sale transactions and ATM withdrawals at high geospatial resolution to gain insight into how people prepare for and recover from disaster.

   (*Source*: Bamberger, M. 2016. *Integrating Big Data into the Monitoring and Evaluation of Development Programmes*. New York: UN Global Pulse. http://unglobalpulse.org/sites/default/files/IntegratingBigData_intoMEDP_web_UNGP.pdf.)

5. Satellite imagery can be used to capture the condition of roofs on houses and develop image processing software to count roofs and identify their construction materials. These data can be ground-truthed against existing survey data related to poverty.

   (*Source*: Bamberger, M. 2016. *Integrating Big Data into the Monitoring and Evaluation of Development Programmes*. New York: UN Global Pulse. http://unglobalpulse.org/sites/default/files/IntegratingBigData_intoMEDP_web_UNGP.pdf.)

6. The Priorities for Local AIDS Control Efforts mapping tool uses a plug-in with QGIS (a popular free and open-source geographic information system program), which allows evaluators to import data (for example, health data) and display them in a simple form on a computer screen and in a printout.

   (*Source*: MEASURE Evaluation. n.d. “The PLACE Mapping Tool: A Plug-in for QGIS.” https://www.measureevaluation.org/resources/tools/hiv-aids/place/the-place-mapping-tool-a-plug-in-for-qgis.)

7. Machine learning and text analytics were used to extract and categorize information from private sector project evaluation reports to create synthetic lessons learned on project performance.

   (*Source*: Bravo, L., A., Hagh, Y. Xiang, and J. Vaessen. Forthcoming. “Machine Learning in Evaluative Synthesis—Lessons from Private-Sector Evaluation in the World Bank Group.” World Bank, Washington, DC.)

8. Machine learning was combined with matching techniques to improve the accuracy of impact estimates in a program evaluation.

(*Source*: Linden, A., and P. R. Yarnold. 2016. "Combining Machine Learning and Matching Techniques to Improve Causal Inference in Program Evaluation." *Journal of Evaluation in Clinical Practice* 22 (6): 864–70.)

9. Machine learning classification algorithms were used to develop predictive models of household poverty status and, combined with satellite imagery, to predict and map poverty across different geographical locations.

   (*Source*: Jean, N., M. Burke, M. Xie, M. Davis, D. B. Lobell, and S. Ermon. n.d. "Combining Satellite Imagery and Machine Learning to Predict Poverty." Sustainability and Artificial Intelligence Lab, Stanford University. http://sustain.stanford.edu/predicting-poverty.)

10. Machine learning algorithms were used to model publicly available survey data and satellite imagery from five African countries (Nigeria, Tanzania, Uganda, Malawi, and Rwanda) and to track and predict local crop yields and economic outcomes.

    (*Source*: You, J., L. Xiaocheng, M. Low, D. B. Lobell, and S. Ermon. n.d. "Combining Remote Sensing Data and Machine Learning to Predict Crop Yield." Sustainability and Artificial Intelligence Lab, Stanford University. http://sustain.stanford.edu/crop-yield-analysis.)

11. Aviation incident reports were categorized using support vector machine analysis, helping to evaluate safety records using archival documents.

    (*Source*: Tanguy, L., N. Tulechki, A. Urieli, E. Hermann, and C. Raynal. 2016. "Natural Language Processing for Aviation Safety Reports: From Classification to Interactive Analysis." *Computers in Industry* 78: 80–95.)

12. Neural networks were used to appraise water quality and flood risks; the resulting data informed management of urban water systems to mitigate water-based disasters.

    (*Source*: Abdellatif, M., W. Atherton, R. Alkhaddar, and Y. Osman. 2015. "Flood Risk Assessment for Urban Water System in a Changing Climate Using Artificial Neural Network." *Natural Hazards* 79 (2): 1059–77.)

13. A combination of machine learning techniques (including support vector machine, K-nearest neighbors, and naïve Bayes) was used to predict

famine onset in Uganda, helping to understand the relationship between food security and starvation.

(*Source*: Okori, Washington, and Joseph Obua. 2011. "Machine Learning Classification Technique for Famine Prediction." *Proceedings of the World Congress on Engineering*, London, July 6–8, 2.)

READINGS AND RESOURCES

Background

Aggarwal, C. C., and C. Zhai, eds. 2012. *Mining Text Data*. New York: Springer Science & Business Media.

Alpaydin, Ethem. 2020. *Introduction to Machine Learning*. Cambridge, MA: MIT Press.

Bamberger, M. 2016. *Integrating Big Data into the Monitoring and Evaluation of Development Programmes*. New York: UN Global Pulse. http://unglobalpulse.org/sites/default/files/IntegratingBigData_intoMEDP_web_UNGP.pdf.

Bamberger, M., and P. York. 2020. *Measuring Results and Impact in the Age of Big Data: The Nexus of Evaluation, Analytics, and Digital Technology*. New York: The Rockefeller Foundation. https://www.rockefellerfoundation.org/wp-content/uploads/Measuring-results-and-impact-in-the-age-of-big-data-by-York-and-Bamberger-March-2020.pdf.

Dayan, P., M. Sahani, and G. Deback. 1999. "Unsupervised Learning." In *The MIT Encyclopedia of the Cognitive Sciences*, edited by R. A. Wilson and F. Keil. Cambridge, MA: MIT Press. http://www.gatsby.ucl.ac.uk/~dayan/papers/dun99b.pdf.

Grimmer, J., and B. M. Stewart. 2013. "Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts." *Political Analysis* 21 (3): 267–97. https://web.stanford.edu/~jgrimmer/tad2.pdf.

Jordan, M. I. 2019. "Artificial Intelligence—The Revolution Hasn't Happened Yet." *Harvard Data Science Review* 1 (1).

Raftree, L., and M. Bamberger. 2014. *Emerging Opportunities: Monitoring and Evaluation in a Tech-Enabled World*. New York: The Rockefeller Foundation. https://www.rockefellerfoundation.org/wp-content/uploads/Monitoring-and-Evaluation-in-a-Tech-Enabled-World.pdf.

Results for Development (R4D) and The International Development Innovation Alliance (IDIA) 2019. "Artificial Intelligence and International Development: An Introduction." R4D, Washington, DC. https://observatoire-ia.ulaval.ca/app/uploads/2019/08/artificial-intelligence-development-an-introduction.pdf.

Advanced

Andersen, Lindsey. 2019. "Artificial Intelligence in International Development: Avoiding Ethical Pitfalls." *Journal of Public and International Affairs* 31. https://jpia.princeton.edu/news/artificial-intelligence-international-development-avoiding-ethical-pitfalls.

Awwad, Y., R. Fletcher, D. Frey, A. Gandhi, M. Najafian, and M. Teodorescu. 2020. *Exploring Fairness in Machine Learning for International Development*. MIT D-Lab | CITE Report. Cambridge: MIT D-Lab. https://d-lab.mit.edu/resources/publications/exploring-fairness-machine-learning-international-development.

Burscher, B., R. Vliegenthart, and C. H. De Vreese. 2015. "Using Supervised Machine Learning to Code Policy Issues: Can Classifiers Generalize across Contexts?" *The Annals of the American Academy of Political and Social Science* 659 (1): 122–31.

Lantz, Brett. 2019. *Machine Learning with R: Expert Techniques for Predictive Modeling*. Birmingham: Packt Publishing.

Meier, P. 2015. *Digital Humanitarians: How Big Data Is Changing the Face of Humanitarian Response*. New York: CRC Press.

Provost, F., and T. Fawcett. 2013. *Data Science for Business: What You Need to Know about Data Mining and Data-Analytic Thinking*. Newton, MA: O'Reilly Media.

Sgaier, Sema K. 2019. "Demystifying Machine Learning for Global Development." *Stanford Social Innovation Review.* Guild, July 24. https://ssir.org/articles/entry/demystifying_machine_learning_for_global_development#.

Other Resources

AidData's GEO program breaks down technological barriers and empowers a broad range of data users to generate analysis and insights with next-generation geospatial data, methods, and tools. https://www.aiddata.org/geo.

Frontier Technologies Hub works with UKAid to apply frontier technologies to the biggest challenges in development. https://medium.com/frontier-technology-livestreaming/.

Geospatial Impact Evaluations rigorously evaluate the impacts and cost-effectiveness of specific development interventions and large investment portfolios with spatial data, leveraging readily available data, like satellite observations or household surveys, to establish a reliable counterfactual for measuring impacts—at a fraction of the time and cost of a "traditional" randomized controlled trial. https://www.aiddata.org/gie.

ICTworks is the premier community for international development professionals committed to using new and emerging technologies to help communities accelerate their social and economic development. https://www.ictworks.org/.

Principles for Digital Development is a website created and maintained by a community of practice of professionals working in digital development. https://digitalprinciples.org.

UK Department for International Development Digital Strategy 2018 to 2020: Doing Development in a Digital World. https://www.gov.uk/government/publications/dfid-digital-strategy-2018-to-2020-doing-development-in-a-digital-world.

USAID's Digital Strategy 2020–2024. https://www.usaid.gov/usaid-digital-strategy.

The following articles discuss similar methodologies:

Hatch, Jonathan. 2019. "6 Ways Artificial Intelligence Is Being Used in International Development." Bond (blog). August 8. https://www.bond.org.uk/news/2019/10/6-ways-artificial-intelligence-is-being-used-in-international-development.

USAID (US Agency for International Development). 2018. "Reflecting the Past, Shaping the Future: Making AI Work for International Development." https://www.usaid.gov/digital-development/machine-learning/AI-ML-in-development.

20 Using Computer-Assisted Qualitative Data Analysis Software

BRIEF DESCRIPTION OF THE METHOD

Over the past 30 years, an increasingly extensive range of software has emerged to analyze and store transcripts from interviews and focus groups, and the use of computer-assisted qualitative data analysis software is now widespread. Available packages facilitate several aspects of qualitative data analysis, including data storage and management, development of coding frameworks, construction of links between codes and coded data, categorization and ordering of codes and coded data, comparison and retrieval of specific coded data, and data display and mapping. All these features can be used to support a more structured and transparent analysis of qualitative data.

THE MAIN VARIATIONS OF THE METHOD

The list of available software platforms for computer-assisted qualitative data analysis is long. The following are widely used:

- *Dedoose* is a web-based tool aimed at facilitating mixed methods research that allows for data coding, analysis, visualization, and basic statistical analyses. The package requires a license.
- *MAXQDA* is a software program developed for the analysis of qualitative and mixed methods data, text, and multimedia materials. It has basic data coding, analysis, and visualization features and is designed to handle large volumes of data. The package requires a license.
- *NVivo* is designed for in-depth analysis of rich text-based and multimedia information. It works with small or large volumes of qualitative data and accommodates both text and images. The package requires a license.
- *Atlas.ti* is a software package for the analysis of unstructured data (text, multimedia, geospatial). It lets the user locate, code, and annotate findings in primary data. It provides analytical and visualization tools to aid interpretation and understanding of complex relations. The package requires a license.

These are just a select few of the many available options.

THE MAIN PROCEDURAL STEPS OF THE METHOD

Given the complexity and diversity of software for computer-assisted qualitative data analysis, no single set of procedural steps can be meaningfully defined.

THE ADVANTAGES AND DISADVANTAGES OF THE METHOD

One advantage of using software for qualitative data analysis is that it allows for more structured and transparent data management and analysis. Many software packages also facilitate identification of differences, commonalities, and relationships among text segments, which provides for a better overview and even understanding of any underlying patterns in the data. Another, perhaps more pragmatic, benefit is that they allow for more efficient data management and analysis, especially when applied on large volumes of text. Finally, their use strengthens the reliability and the potential for replicability of the analysis. The systematic treatment of data facilitates triangulation among data, which may strengthen the validity of findings.

One shortcoming is that software for qualitative analysis is sometimes mistaken for an analytical strategy or is thought of as performing the data analysis itself. This is an unfortunate misconception. Although software packages may facilitate or support qualitative analysis, the coding and interpretation still rest with the evaluator. Another issue worth considering is that preparing and entering data may be time-consuming, depending on the type of data and the software chosen.

THE APPLICABILITY OF THE METHOD

Software for qualitative data analysis can be applied to any type of qualitative data, including transcripts from individual interviews, focus groups, or open-ended questions in surveys. There are also software packages available for the analysis of imagery and audio data. Examples include the following:

1. MAXQDA, a content analysis software, was used for managing and coding transcripts from 78 interviews as part of an evaluation of how organizational incentives, norms, culture, and practices shape the production and use of self-evaluations in the Bank Group.

   (*Source*: World Bank. 2016. *Behind the Mirror: A Report on the Self-Evaluation Systems of the World Bank Group*. Washington, DC: World

Bank. http://ieg.worldbank.org/sites/default/files/Data/Evaluation/files/behindthemirror_0716.pdf.)

2. Two software packages, Sonar Professional and Atlas.ti, were used in a portfolio review of approaches to social accountability.

   (*Source*: Ringold, Dena, Alaka Holla, Margaret Koziol, and Santhosh Srinivasan. 2011. "Portfolio Review Methodology." In *Citizens and Service Delivery: Assessing the Use of Social Accountability Approaches in Human Development Sectors*. World Bank's Directions in Development—Human Development. Washington, DC: World Bank.)

3. MaxQDA was used to transcribe, code, and analyze data from 56 focus groups and additional key informant interviews as part of an evaluation of the World Bank's nutrition programs in Malawi.

   (*Source*: Osendarp, Saskia Josepha Maria, Forhad J. Shilpi, Timothy Gondwe, Innocent Pangapanga-Phiri, Alexander Kalimbira, Beatrice Mtimuni, Deusdedit Kafere, Gabriella Chuitsi, Felix Phiri, and Ziauddin Hyder. 2019. "Determinants of Reductions in Childhood Stunting in Malawi's Community-Based Nutrition Programs." Discussion Paper, Health, Nutrition, and Population, World Bank Group, Washington, DC. http://documents.worldbank.org/curated/en/297601565964816621/Determinants-of-Reductions-in-Childhood-Stunting-n-Malawis-Community-based-Nutrition-Programs.)

4. An evaluation of the World Bank's support to national statistical offices used NVivo to code and analyze the content of 76 semistructured interviews with World Bank staff and external experts.

   (*Source*: World Bank. 2018. *Data for Development: An Evaluation of World Bank Support for Data and Statistical Capacity*. Independent Evaluation Group. Washington, DC: World Bank. http://ieg.worldbankgroup.org/evaluations/data-for-development.)

5. Dedoose was used to code and analyze transcripts from focus groups in a qualitative analysis of caregiver perceptions of children's linear growth in Bangladesh.

   (*Source*: Hossain, M., S. Ickes, L. Rice, G. Ritter, N. Naila, T. Zia, B. Nahar, M. Mahfuz, D. M. Denno, T. Ahmed, and J. Walson. 2018. "Caregiver

Perceptions of Children's Linear Growth in Bangladesh: A Qualitative Analysis." *Public Health Nutrition* 21 (10): 1800–1809.)

6. NVivo was used to analyze interviews with 96 stakeholders to evaluate the contribution of mobile phones to rural livelihoods and poverty reduction in Tanzania.

   (*Source*: Sife, Alfred Said, Elizabeth Kiondo, and Joyce G. Lyimo-Macha. 2010. "Contribution of Mobile Phones to Rural Livelihoods and Poverty Reduction in Morogoro Region, Tanzania." *The Electronic Journal of Information Systems in Developing Countries* 42 (1): 1–15.)

7. Atlas.ti was used to code and analyze 2,000 pages of interview transcripts as part of a review of participatory water management in Bangladesh.

   (*Source*: Dewan, C., M-. C. Buisson, and A. Mukherji. 2014. "The Imposition of Participation? The Case of Participatory Water Management in Coastal Bangladesh." *Water Alternatives* 7 (2): 342–66. http://www.water-alternatives.org/index.php/all-abs/250-a7-2-4/file.)

8. MaxQDA was used to process qualitative data from focus group discussions and cluster them based on keywords in a study of local perceptions of climate change in Indonesia.

   (*Source*: Boissière, M., B. Locatelli, D. Sheil, M. Padmanaba, and E. Sadjudin. 2013. "Local Perceptions of Climate Variability and Change in Tropical Forests of Papua, Indonesia." *Ecology and Society* 18 (4). www.jstor.org/stable/26269394.)

READINGS AND RESOURCES

Background

Friese, Susanne. 2019. *Qualitative Data Analysis with ATLAS.ti*. Thousand Oaks, CA: SAGE.

Jackson, Kristi, and Pat Bazeley. 2019. *Qualitative Data Analysis with NVivo*, 3rd ed. Thousand Oaks, CA: SAGE.

Kuckartz, Udo, and Stefan Rädiker. 2019. *Analyzing Qualitative Data with MAXQDA: Text, Audio, and Video*. Cham, Switzerland: Springer. https://link.springer.com/content/pdf/10.1007/978-3-030-15671-8.pdf.

Salmona, Michelle, Eli Lieber, and Dan Kaczynski. 2019. *Qualitative and Mixed Methods Data Analysis Using Dedoose: A Practical Approach for Research Across the Social Sciences*. Thousand Oaks, CA: SAGE.

Silver, C., and A. Lewins. 2014. *Using Software in Qualitative Research*. Thousand Oaks, CA: SAGE.

Advanced

Hutchison, Andrew, Lynne Johnston, and Jeff Breckon. 2010. "Using QSR-NVivo to Facilitate the Development of a Grounded Theory Project: An Account of a Worked Example." *International Journal of Social Research Methodology* 13: 283–302.

Leech, N. L., and A. J. Onwuegbuzie. 2011. "Beyond Constant Comparison Qualitative Data Analysis: Using NVivo." *School Psychology Quarterly* 26 (1): 70–84.

Lewis, R. B. 2004. "NVivo 2.0 and ATLAS.ti 5.0: A Comparative Review of Two Popular Qualitative Data-Analysis Programs." *Field Methods* 16 (4): 439–64.

Paulus, Trena, and Jessica Lester. 2016. "ATLAS.ti for Conversation and Discourse Analysis Studies." *International Journal of Social Research Methodology* 19 (4): 405–28.

Saillard, E. 2011. "Systematic Versus Interpretive Analysis with Two CAQDAS Packages: NVivo and MAXQDA." *Forum Qualitative Sozialforschung / Forum: Qualitative Social Research* 12 (1), article 34.

APPENDIX A

GLOSSARY OF KEY TERMS

The purpose of the glossary is to define key terms used in this guide. To the extent possible, definitions from the glossary of key evaluation terms of the Development Assistance Committee of the Organisation for Economic Co-operation and Development (OECD-DAC; http://www.oecd.org/dac/) or from Independent Evaluation Group publications are provided.

Activity. Actions taken or work performed through which inputs, such as funds, technical assistance, and other types of resources are mobilized to produce specific outputs (OECD-DAC).

Alternative explanation. A plausible or reasonable explanation for changes in an outcome variable caused by factors other than the program under evaluation.

Analytical generalization. A nonstatistical approach for generalization of findings based on a theoretical analysis of the program and contextual factors producing outcomes.

Analytical technique. An approach used to process and interpret information as part of an evaluation (OECD-DAC).

Assumption. A hypothesis about factors or risks that could affect the progress or success of a development intervention (OECD-DAC).

Attribution. The ascription of a causal link between changes observed or expected to be observed and a specific intervention (OECD-DAC).

Baseline. A measure describing the situation before a development intervention, against which progress can be assessed or comparisons made (OECD-DAC).

Bayesian updating. A technique for refining the probability that a hypothesis or theory is true (or false) as more information becomes available.

Big data. Data characterized by high volume, velocity (real time), and variety (wide range of information).

Causal description. Determining the outcome(s) attributable to a program.

Causal explanation. Clarifying the mechanisms through which a program generates the outcome(s).

Comparison/control group. The group of individuals in an experiment (control) or quasi-experiment (comparison) who do not receive the treatment program.

Contribution. A program effect that is difficult to isolate from other co-occurring causal factors.

Counterfactual. The situation or condition that may have hypothetically materialized for individuals, organizations, or groups if no development intervention had been implemented.

Data analytics. An umbrella term for analytical techniques and processes used to extract information from data, including data collected with emerging technologies and big data (see definition above).

Data collection method. An approach used to identify information sources and collect information during an evaluation (OECD-DAC).

Discount rate. The interest rate used in cost-benefit analysis to adjust the value of past or future cash flows to present net value.

Doubly decisive test. A type of test (made famous by process tracing literature) that is both strong and symmetrical; that is, it can either strengthen or weaken the hypothesis considerably, depending on whether the test is positive or negative, respectively.

Effect. An intended or unintended change attributable directly or indirectly to an intervention (OECD-DAC).

Effect size. A quantitative measure of the outcome difference between a treatment group and a comparison/control group.

Effectiveness. The extent to which the development intervention's objectives were or are expected to be achieved, taking into account their relative importance (OECD-DAC).

Efficiency. A measure of how economically resources/inputs (funds, expertise, time, and so on) are converted to results (OECD-DAC).

Ex ante evaluation (also known as prospective evaluation). An evaluation that is performed before implementation of a development intervention (OECD-DAC).

Ex post evaluation (also known as retrospective evaluation). Evaluation of a development intervention after it has been completed (OECD-DAC).

External validity. The extent to which findings from an evaluation can be generalized to other, perhaps broader, settings and groups.

Evaluation theory. Approaches to evaluation that prescribe a specific role and purpose for the evaluation.

Hoop test. A type of test (made famous by process tracing literature) that is strong but not symmetrical: it can substantially weaken the theory or hypothesis if negative but cannot strengthen it if positive.

Impact. Positive and negative, primary and secondary long-term effects produced by a development intervention, directly or indirectly, intended or unintended (OECD-DAC).

Independent evaluation. An evaluation carried out by entities and persons free of the control of those responsible for the design and implementation of the development intervention (OECD-DAC).

Influencing factor. An aspect of the program implementation context that affects the program outcomes, either qualitatively or quantitatively.

Input. The financial, human, and material resources used for the development intervention (OECD-DAC).

Internal validity. The credibility or "truth value" of a causal connection between a program and specific outcome.

Judgmental matching. The use of nonstatistical techniques to establish comparable treatment and comparison groups for the purposes of net effect estimation.

Logic model. A depiction (often in tabular form) of the activities, outputs, and outcomes of a program. The term is often used interchangeably with *program theory*. Many logic models differ from program theories in that they merely list program activities, outputs, and outcomes instead of explaining how they are connected.

Logical framework (logframe). A management tool used to improve the design of interventions, most often at the project level. It involves identifying strategic elements (inputs, outputs, outcomes, impact) and their causal relationships, indicators, and the assumptions or risks that may influence success and failure. It thus facilitates planning, execution, and evaluation of a development intervention (OECD-DAC).

Mechanism. The underlying processes generating an outcome.

Outcome. The likely or achieved short-term and medium-term effects of an intervention's outputs (OECD-DAC). Outcomes are usually in the form of behavioral or organizational changes.

Output. The products, capital goods, and services that result from a development intervention; it may also include changes resulting from the intervention that are relevant to the achievement of outcomes (OECD-DAC).

Pipeline approach. A technique for comparison group selection, where the comparison group is composed of individuals who have been selected (eligible) to participate but have not (yet) been involved or benefited from intervention activities.

Program. A set of activities and outputs intended to advance positive outcomes for a specific group of people, here used as generic term for a policy intervention.

Program theory. A visual and narrative description of how the activities and outputs of a program are expected to generate one or more outcomes.

Propensity score matching. The use of statistical (regression-based) techniques to establish a comparison group that is equivalent to the treatment group for purposes of net effect estimation.

Power (statistical). The probability that a statistical test (based on a sample) will detect differences (when they truly exist in the population).

Purposive sampling. A nonrandom sampling procedure.

Random allocation. The random selection of participants to the treatment group and the comparison group, whereby selection bias from observed and unobserved characteristics is eliminated.

Random sample. A sample drawn from a population where each unit has an equal probability of being selected.

Regression. A statistical procedure for predicting the values of a dependent variable based on the values of one or more independent variables.

Reliability. Consistency or dependability of data and evaluation judgments, with reference to the quality of the instruments, procedures, and analyses used to collect and interpret evaluation data (OECD-DAC).

Sample. A subset of units (for example, individuals or households) drawn from a larger population of interest.

Sampling. A process by which units (for example, individuals or households) are drawn from a larger population of interest. See also *random and purposive sampling procedures*.

Selection bias. Bias introduced when specific individuals or groups tend to take part in the program (treatment group) more than other groups, resulting in a treatment group that is systematically different from the control group.

Sensitivity analysis. Determines how sensitive the findings are to changes in the data sources or the data collection and analysis procedures.

Smoking gun test. A type of test (made famous by process tracing literature) that is strong but not symmetrical; it can considerably strengthen the theory or hypothesis if positive but cannot weaken it if negative.

Stock and flow diagram. A visual depiction of causal relationships in a system modeled on the basis of one or more stocks (for example, the total number of rural farmers under the poverty line) and the flows between them that change the stock values (for example, job growth, currency inflation).

Straw-in-the-wind test. A type of test (made famous by process tracing literature) that is both weak and symmetrical; it can never substantially strengthen nor weaken the theory or hypothesis.

APPENDIX B

DEVELOPING PROGRAM THEORIES

Program theories,[1] also referred to as logic models or theories of change (and other slightly different terms), are widely used in evaluation. A program theory can be broadly defined as a visual and narrative description of the main program inputs, activities, outputs, and desired outcomes. A central aspect of a program theory is the specification of how these are connected, that is, how the program activities and outputs are assumed to generate the desired outcomes. Program theories are now commonly required by development agencies as part of project planning.

There are many different ways of developing program theories as part of an evaluation. They may be developed prospectively, in the early phase of program design, or retroactively, after the program has entered the implementation phase. In some cases, the evaluator develops the program theory on the basis of program documents, interviews with program staff, or some combination of these. In other cases, program theory development is a collaborative effort between the evaluator and program staff, and perhaps including other stakeholders. These collaborative efforts can be structured around one or more workshops. Finally, program theories may be informed by existing research, relevant social scientific theories, and past evaluations.

There are several reasons for the widespread use of program theories in evaluation. First and foremost, program theories allow for a shared understanding between the evaluation team and program staff of how and in what way the program is intended to bring about change. This shared understanding of how the program is intended to function is important because (among other things) it may improve collaboration, foster agreement on evaluation findings, or reduce tensions. Second, program theories are often tested empirically in the course of evaluations and, as such, they focus the design of the data collection process. Third, a well-developed and well-tested program theory is an essential part of the lessons learned through an evaluation because it facilitates a deeper understanding of how and why the program worked or failed to work. This type of information is essential to inform future planning and program design.

Using program theories in evaluation has plenty of benefits but also presents a number of challenges. One common challenge is poor conceptual framing: how program components are causally connected among themselves and with the desired outcomes is often not well enough detailed, referenced, or specified, and the causal links are either unconvincing or omitted.

Another common issue emerges from the typical disconnect between the program theory and the data collection process; although the former should drive the latter, in practice some parts of theories are often untestable, and confidence in their veracity can be neither strengthened nor weakened with rigorous procedures. A related problem is construct validity: the development of measurements and indicators is often poorly linked with program theory.

Finally, program theories are prone to confirmation bias: the discussion of influencing factors and alternative explanations is often poor or altogether omitted. As a result, many program theories are overly abstract or simplistic and fail to support any in-depth or defensible examination and understanding of how the program works and brings about change under certain circumstances.

Despite these challenges, in situations where the emphasis is on understanding and explaining why and how a program brings about change, it is essentially impossible to avoid dealing with program theories. Therefore we propose a checklist of minimum requirements that program theories should fulfill. To realize their potential and add value to the evaluation, program theories should ideally contain the following elements:

1. Identify all the program activities, outputs, and intermediate outcomes that are essential to understand the causal logic of how a program works and brings about change;

2. Explain in sufficient detail how and why these parts are connected;

3. Specify the external influencing factors (contextual conditions, other programs, and other processes and activities) that could affect program implementation, delivery, and outcomes;

4. Clearly distinguish (and potentially choose) between theory of action (focused on causal linkages between implementation and delivery) and theory of impact (focused on causal linkages between delivery and outcomes), which allows for a distinction between implementation failure and theory failure;[2] and

5. To the extent possible, formulate alternative explanations (rival hypotheses) that might have produced changes in the program outcomes.

For program theory to be fruitfully used and integrated into the evaluation, it should inform the design of data collection and data analysis. In particular, the evaluator should do the following:

1. Ensure that data collection covers the most salient program activities, outputs, and outcomes (as detailed in the program theory) and pay attention to both intended and unintended outcomes (positive and negative);
2. Ensure that data collection covers the most salient alternative explanations and influencing factors;
3. Examine how the collected data support or bring into question specific aspects of the program theory; and
4. Refine and modify the program theory as informed by the data.

Together, these guidelines should facilitate a more productive use of program theories in evaluation.

Readings

Astbury, B., and F. L. Leeuw. 2010. “Unpacking Black Boxes: Mechanisms and Theory Building in Evaluation.” *American Journal of Evaluation* 31 (3): 363–81.

Bickman, Leonard, ed. 1987. “Using Program Theory in Evaluation.” Special issue, *New Directions for Program Evaluation* 1987 (33).

Brousselle, Astrid, and François Champagne. 2011. “Program Theory Evaluation: Logic Analysis.” *Evaluation and Program Planning* 34 (1): 69–78.

Funnel, S. C., and P. J. Rogers. 2011. *Purposeful Program Theory—Effective Use of Theories of Change and Logic Models*. San Francisco: Jossey-Bass.

Leeuw, F. L. 2003. “Reconstructing Program Theories: Methods Available and Problems to Be Solved.” *American Journal of Evaluation* 24 (1): 5–20.

Leroy, Jef L., Marie Ruel, and Ellen Verhofstadt 2009. “The Impact of Conditional Cash Transfer Programmes on Child Nutrition: A Review of Evidence Using a Programme Theory Framework.” *Journal of Development Effectiveness* 1 (2): 103–29.

Petrosino, Anthony, Patricia J. Rogers, Tracy A. Huebner, and Timothy A. Hacsi, eds. 2000. “Program Theory in Evaluation: Challenges and Opportunities.” Special issue, *New Directions for Program Evaluation* 2000 (87).

Rogers, P. J. 2000. “Program Theory: Not Whether Programs Work but How They Work.” In *Evaluation Models*, edited by D. L. Stufflebeam, G. F. Madaus, and T. Kellaghan, 209–32. Evaluation in Education and Human Services vol. 49. Dordrecht: Springer.

W. K. Kellogg Foundation. 2006. *Logic Model Development Guide*. Battle Creek, MI: W. K. Kellogg Foundation.

Weiss, C. H. 1997. "Theory-Based Evaluation: Past, Present, and Future." *New Directions for Evaluatio*n 1997: 41–55.

Endnotes

[1] The term *program* is used in a generic sense to refer to any type of policy intervention (activity, project, program, policy, and so on). One could use the term *intervention theory* instead of the better-known term *program theory*.

[2] Failure for outcomes to emerge can either be due to implementation failure (the program outputs were not delivered) or theory failure (the program outputs were delivered but did not make a difference—that is, they may not have been the right solution to the problem in the given circumstances), or a combination of both.

ABOUT THE AUTHORS

Jos Vaessen, PhD, has been an adviser on evaluation methods at the Independent Evaluation Group of the World Bank Group since 2016. Since 1998 Jos Vaessen has been involved in evaluation research activities, first as an academic and consultant to bilateral and multilateral development organizations, and from 2011 to 2015 as an evaluation manager at UNESCO. Jos is a member of the Evaluation Advisory Panel at UNICEF and regularly serves on reference groups of evaluations for different institutions.

Sebastian Lemire, PhD, is an associate at Abt Associates. He brings over 15 years of experience designing and managing large-scale, multiyear evaluations in the fields of education, social welfare, and international development. His publications center on innovative approaches and analytical techniques for theory-based impact evaluation. Sebastian currently serves on the Editorial Advisory Board of *Evaluation* and as an associate editor for the *American Journal of Evaluatio*n.

Barbara Befani, PhD, is an evaluation methods specialist with 15 years of experience in research, training, and advisory work. She specializes in hybrid (quali-quanti) methods for impact evaluation; the theory and practice of methods choice; and unified approaches to evaluation quality (that apply to qualitative, quantitative, and mixed method designs). She is or has been an associate of the Institute for Development Studies at the University of Sussex, the University of East Anglia, and the University of Surrey.

IEG
Independent Evaluation Group
WORLD BANK GROUP
World Bank • IFC • MIGA

The World Bank
1818 H Street NW
Washington, DC 20433