11 Policy evaluation and behavioral economics

Kai Ruggeri, Julia P. Stuhlreyer, Johanna Emilia Immonen, Silvana Mareva, Maja Friedemann, Alessandro F. Paul, Matthew Lee, and Rachel C. Shelton

Acknowledgments: Annalisa Robbiani, Frederick W. Thielen, Amiran Gelashvili, Filippo Cavassini, and Faisal Naru

Chapter summary

Empirical policy evaluation is necessary to answer the question: what is a good policy? To answer this question, this chapter will review the management of policy evaluation and cover key performance indicators for evaluation, such as efficiency and fidelity. On the basis of these indicators, common frameworks for policy evaluation are explained. Some of the challenges in conducting policy evaluation, though, are the complex and variable aspects of all policies, as well as their context-specific antecedents and outcomes. This makes comparison between policies very challenging, if not impossible; hence, we close with an example of strategies for standardizing the evaluation of policies across domains and approaches. Ultimately, the purpose of this chapter is to identify not only what comprises a strong policy but also *how* to produce effective policies that maximize the number of people reached and impacted.

Learning objectives

- Understand the important aspects and necessary steps of policy evaluation
- Distinguish key policy performance indicators: efficacy, effectiveness, efficiency, fidelity, adaptation, and sustainability
- Recognize the main elements of common approaches to evaluate policy
- Be able to develop a general policy evaluation framework
- Utilize a standardized scoring system for evaluating different types of policies

Introduction

In the second chapter of this book, we covered processes of policy development and implementation, particularly as they relate to evidence. In these processes, the ultimate aim is to generate the best possible outcomes for the lowest costs in resources. Clear structures to

DOI: 10.4324/9781003181873-11

implement, manage, and track policies are paramount for establishing accountability among stakeholders and facilitating later analysis (see, for example, Ministry of Foreign Affairs, 2009; Australian Development Agency, 2009). However, no matter how much we invest in systematic structures or uses of evidence to inform the policy content, we cannot know if a policy is effective unless we perform an appropriate evaluation.

There are many possible ways to evaluate a policy. Besides being based on scientifically rigorous evidence, Lindblom (1959), for instance, classically suggested that a "good" policy is one that achieves agreement about a particular desired outcome across key decision-makers. While this may suffice on some superficial levels, a more robust and comprehensive assessment of the benefits and shortcomings of policies can be obtained through rigorous **evaluation**, which entails generating information on expected and actual impacts of specific policies, as well as the processes involved in their development, implementation, sustainability, and use of resources (OECD, 2015a). Evaluation is also to be distinguished from **monitoring**, which refers to the continuous assessment of implementation against an agreed-upon schedule (OECD, 2015a).

To explain policy evaluation, the chapter is structured as follows – first, we discuss policy implementation and how evaluation is embedded within it. On this basis, we provide a coarse **framework** (an overarching guideline with principles for a construct that can help direct the approach taken without specifying or itemizing all content) detailing what should be established in an evaluation process and how it could be managed. Subsequently, common approaches to policy evaluation are presented in a simplified framework with special emphasis on key indicators, such as efficacy, fidelity, adaption, and sustainability. Further, the chapter reviews how data could be collected and interpreted to assess whether a given policy is a "good" policy beyond Lindblom's classic description. Finally, we present a systematic scoring system that allows the comparison of different policies across topics and contexts and their present and potential uses.

Beginning with the journey in mind

The degree to which the insights generated by evaluations are useful strongly depends on the quality and type of data and evidence gathered: claims about a policy's effects on an attribute of interest would only be appropriate if relevant supportive data were obtained. Ideally, to enhance generalizability and external validity, the data should be obtained across different types of contexts and populations. It is good practice to determine and define the criteria for interpreting whether a policy has met its goal or reached the desired outcome a priori. Additionally, in order to promote transparency and maintain accountability, these criteria can be determined in collaboration with policymakers and other key stakeholders and then be shared publicly (e.g. using an **open data framework**) (Zuiderwijk & Janssen, 2014). Retrospectively applied effectiveness criteria are often chosen arbitrarily and are easily disputable. The observation that a decision had a beneficial outcome does not readily mean that the decision was good or that the outcome was planned. Therefore, plans for impact or outcome assessment should have already been considered and determined as the selection of the most suitable policy for a given issue is made. To facilitate policy design and planning, two common approaches are used: ***ex ante*** (before the policy implementation) and ***ex post*** impact (after the policy implementation) (OECD, 2014a).

Ex ante *evaluation*

Ex ante assessment focuses on the planning and design of the policy itself and poses the question, "Will this policy have an impact and, if so, in what way?" To do this, *ex ante* also has to consider "What is the problem that needs to be resolved?" as well as "What are the likely intended and unintended outcomes of this policy?" The *ex ante* assessment is useful to establish the need for a policy prior to the actual implementation. An *ex ante* assessment provides the advantage of choosing among different policy options on the basis of expected impacts – the flexibility in changing implementation strategies, through planned and well-documented adaptation, comes from the implementation design and real-world practice. The implementation design includes the means through which the policy objectives are realized as well as the objectives themselves. An example of an approach to *ex ante* impact assessment is presented in Box 11.1.

BOX 11.1 EUROPEAN COMMISSION GUIDELINES FOR POLICY PLANNING

There is no "gold standard" for how policies ought to be planned and designed, and the process may differ considerably across organizations, contexts, and levels of policymaking (e.g. local, state, national). Nevertheless, it may be valuable for the various stakeholders working in different segments of the policy cycle to have a schematic understanding of how such a process may be applied in practice. As an example, the guidelines for *ex ante* policy planning applied by the European Commission (2016) are presented next.

1 Bring together an inter-service group consisting of people who work in fields related to the subject that will be evaluated
2 Inter-service group prepares the impact assessment
3 Announce to stakeholders and policymakers that they can provide feedback about the potential challenges and the impact of the implemented policy
4 Commence a 12-week, open public consultation to make sure that the stakeholders and policymakers have the chance to voice their opinions
5 Collection of data, input from stakeholders, and further evidence
6 Write the impact assessment report
7 Send report to the Regulatory Scrutiny Board for review; the review includes a closer look at what can be improved and formulates advice for future policies
8 If the report is accepted by the board, submit further policy initiatives to inter-service consultation

Policy implementation

After considering these many factors, the move toward systematically realizing policies can begin. It is important to establish the implementation plan early to avoid potential biases as well as structural barriers (financial, organizational, transactional) that can lead to bottlenecks and other challenges later in the process. The policy implementation plan translates an idea

into policy – it details how a policy is put into action, defines the monitoring process, and ensures that all planned aspects are performed (Centers for Disease Control and Prevention, 2015). Implementation plans comprise a set of selected implementation strategies intended to achieve desired outcomes (e.g. enhancing the speed and quality at which a policy is adopted, implemented, and sustained). For example, implementation strategies include steps such as accessing new funding, starting a dissemination organization, or conducting educational meetings (Powell et al., 2015). Further, the implementation plan should contain clear guidelines defining at which point a policy is considered fully implemented. In practice, policies often need to be rolled out incrementally over time and space; thus, it is useful to distinguish between initial, medium, and longer term policy implementation when designing a policy evaluation. For example, some policies are immediately ready for enforcement and monitoring, whereas more complex policies (e.g. the Paris Agreement on climate change) require several years to achieve full implementation. Similarly, some areas within a jurisdiction are immediately ready to implement a new policy, whereas other areas require additional capacity building first. The knowledge about how an intervention was executed provides the basis for assessing its effectiveness and impact. The information about what policy aspects were implemented and the extent to which they were achieved allows us to evaluate how the policy components that were originally planned relate to the observed effects. Furthermore, an understanding of these factors allows policy evaluators to recognize when and to what extent a policy is not being implemented as planned (e.g. with low fidelity) which can lead to the termination of potentially effective policies and the continuation of ineffectively implemented policies (Brownson et al., 2015). Knowledge of how to adapt policies that are being mis-implemented to get them back on track and even when to de-implement (e.g. replace, terminate, or defund) ineffective and potentially harmful policies are key actions that all policy evaluators must consider.

Ex post *evaluation*

To assess whether a policy has been implemented appropriately and effectively, robust assessment is critical (Howlett et al., 2015). This is known as *ex post* policy evaluation, which is a systematic assessment of progress made toward meeting objectives, implementation processes, and the integration of relevant evidence and methods (World Health Organization, 2007). An *ex post* assessment explores whether a policy's impact goal was reached and, thus, determines the effectiveness of an intervention and the need for any alternative action (for example, revising or adapting existing policies and adjusting the implementation plan). Without evaluation, it is impossible to determine effectiveness and whether further adoption and wider policy dissemination are appropriate. The World Health Organization (WHO) divides policy evaluation into two forms. First, the evaluation focuses on the content of the policy – its vision, objectives, and target areas. Second, the evaluation of the plan refers to assessing the proposed implementation strategies, targets, and their indicators. Furthermore, because there are numerous possible ways to approach an evaluation, every evaluation team can choose the procedure that is most suitable to a given policy and its implementation (Trochim, 2009). It is important to note that evaluation is not a static process, but it can be iterative and flexibly adjusted according to the context (Menon, Karl, & Wignaraja, 2009).

Process

According to the WHO (2007), it is critical to evaluate the full scope of the development process of a policy. This includes considerations of whether the development followed the

best practice guidelines, whether the key stakeholders were involved in the process, and the extent to which the plan was checked against the available resources and the best available evidence. Further, a strong policy plan would take local conditions and needs into account when detailing key areas of action and steps necessary for successful implementation by specific stakeholders (Menon et al., 2009). The evaluation assesses whether these actions are taken, and in instances where this is not the case, it analyzes the reasons behind the failure to adhere to the plan. Evaluation, therefore, may inform all aspects of a given policy and may provide insights into iterations of related initiatives, regardless of whether the planned outcomes are met or not. For this reason, the United Nations Development Programme (Menon et al., 2009) considers policy planning, monitoring, and evaluation to be interrelated processes, which provides an important link between past, present, and future initiatives and development results. Equally important is the need to evaluate the quality of the regulatory and policy tools that are being used to develop the policy (OECD, 2014b).

Monitoring

Have you ever made a New Year's resolution? If you are like many people, you probably committed to changing something, but shortly after January 1 the effort going into those changes started tapering away, and achieving the goals you set out became increasingly unlikely. But how early were you aware that things were not going as planned, and what could you have changed had you known? This is not so different from monitoring policies, which is often a critical feature of interventions during and after implementation.

Monitoring is a key element in evaluations because there are numerous factors, such as changes in the environment or staff, that make policy implementation challenging (WHO, 2007). This is important given that the contexts and settings in which policy implementation is taking place are complex and dynamic. Monitoring is defined as the ongoing process of assessing the progress of the policy implementation strategies toward the set goal (Menon et al., 2009; WHO, 2007). In this regard, it is necessary to determine whether the pre-specified actions were performed, if the progress proceeded as planned, and whether any difficulties or unanticipated challenges arose (Menon et al., 2009; WHO, 2007). On this basis, the planned and performed implementation strategies can be adjusted to ensure that the desired goals and policy objectives are successfully met (Menon et al., 2009). The resulting findings may be used to communicate and engage with the key stakeholders about the status and the advances made toward the objectives (Menon et al., 2009).

There is no fixed timeline for when monitoring should begin or how long it should last. However, Waterman and Wood (1993) outline the four key stages of monitoring, each being an iteration of the previous one:

Stage 1 – Collect the facts: examine qualitative information to understand if a policy-related issue exists
Stage 2 – Identify relevant stimuli: produce a database of information for regular evaluation
Stage 3 – Statistical analysis: evaluate changes over time to understand impacts
Stage 4 – Re-examine initial information: use all insights once policy is fully implemented to go back to policymakers to understand and consider further actions

In some cases, these approaches may be outdated in modern policy contexts, yet the general nature of the timeline remains relevant.

The main contribution of monitoring is that it identifies adjustments needed within implementation of the policy, particularly when the data indicate problems and the need for redirecting course. When this occurs, "patches" can address the identified issues and help to avoid repeating mistakes (Howlett et al., 2015). As policies usually comprise bundles of multiple policy tools, instead of abandoning existing policies and replacing them with new ones, policy designers can restructure policies by adding or subtracting elements or objectives to or from the existing policy mix. Of course, the specific context and likely results of the policy redesign need to be taken into account when introducing changes over time (Howlett & Rayner, 2013).

Monitoring itself should not replace a full evaluation, but the information that it produces may be a useful tool in terms of transparency (i.e. ad hoc reporting on progress to stakeholders) and overall effectiveness (by allowing one to identify and correct issues earlier in the process). While monitoring is different from evaluation, monitoring activities often produce the data needed to inform and drive key evaluation activities.

Evaluation management

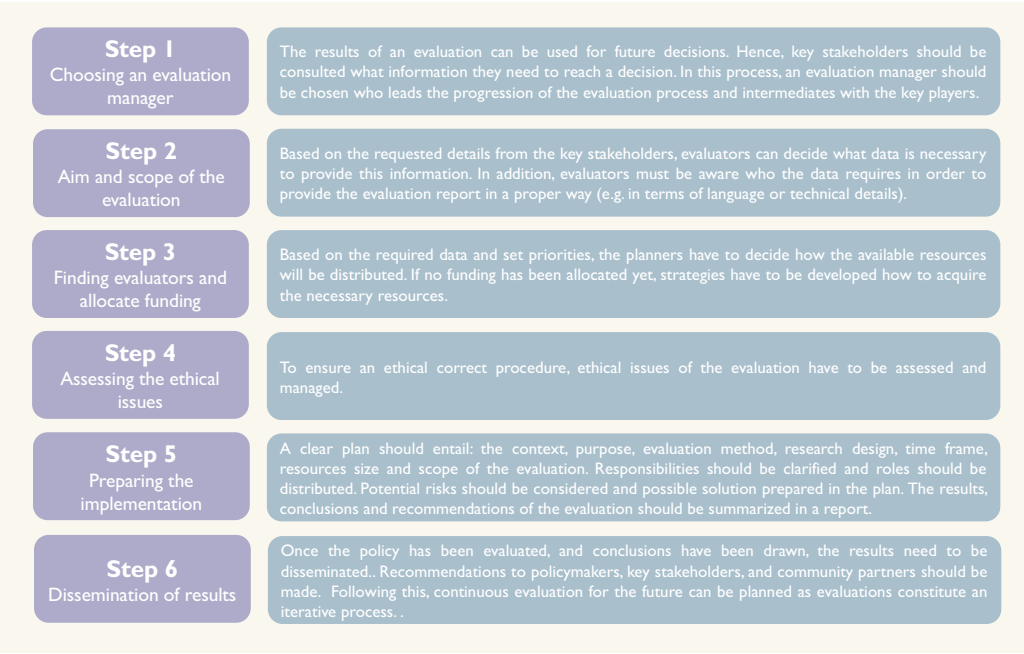
Evaluations are often complex processes that surpass the workload capacity of individuals and instead require full teams. This includes the input and engagement of stakeholders and partners, as they may provide feedback and contribute to decision-making across every step of the evaluation (Menon et al., 2009). Roles and responsibilities should be attributed on the basis of competencies and, when necessary, external expertise (e.g. from statisticians, analysts) should be contracted. Additionally, an evaluation team should oversee the whole evaluation process and should compile the information acquired by the different members of the evaluation group. The evaluation team is also responsible for maintaining communication with, and disseminating key findings to, policymakers, key stakeholders, and community partners.

The European Commission (EC; European Commission, 2016) states that to evaluate policies effectively, the evaluation team should fulfill the following features: political support, resources, expertise, coverage, integration, and structure. To plan and perform the evaluation process, the evaluation group and defined stakeholders should designate an evaluation manager, who intermediates among the involved parties and takes on the daily responsibilities of the evaluation (Menon et al., 2009). According to the United Nations Development Program (UNDP) (Menon et al., 2009), the evaluation manager should lead the evaluation process, debrief the evaluators, coordinate the different parties involved, provide the relevant data to the key stakeholders, manage the contract agreements, and review the evaluation plan and reports. Following the evaluation, a management response, consisting of a statement about recommendations and further procedures, should be written and monitored until all planned actions are taken or canceled (Fertman & Allensworth, 2016; Rogers, 2014).

There are many different approaches to performing an evaluation. The choice of which approach to use is based on the questions one is trying to address, the feasibility and resources available, the criteria that will be used to judge program performance, and the performance standards that must be reached for the program to be considered successful (Community Tool Box, n.d.). Although there is no single "just" approach to evaluations, similar considerations apply to most planned and applied evaluations (see Figure 11.1).

The following steps outline practical elements within an adapted evaluation framework of the WHO (WHO, 2007) and the UNDP (Menon et al., 2009). Evaluations are complex

FIGURE 11.1 A pragmatic framework for evaluation and monitoring



processes that require numerous considerations beyond the ones mentioned in the oversimplified version outlined here. Nonetheless, the framework provides a useful basis for evaluation planning.

Features of evaluation

The literature on policy evaluation converges on several features (see Table 11.1) considered to be **key performance indicators (KPIs)** in any policy evaluation. The features are applied and explained in an example concerning a municipality that aims to reduce the use of personal cars to reduce traffic congestion by distributing free public transport tickets.

When one evaluates the sustainability of a policy in the long term, it is necessary to establish that the effectiveness is not a mere **superficial effect** in which the expected results are initially shown but do not last (Loewenstein & Chater, 2017; Sunstein, 2017a). Consider an example from the New York City (NYC) Human Resources Administration Department of Social Services (HRA-DSS), which is the largest US social services agency combating poverty and income inequality. For decades, HRA-DSS workforce development programs tried to help NYC residents to move out of poverty and off welfare. This target was pursued by placing individuals into jobs as quickly as possible, without much consideration of the sustainability of this impact or unintended consequences. This approach inadvertently resulted in individuals not being trained in line with long-term employability goals. This placed them in lower wage positions with little access to higher wage sectors. The resulting

TABLE 11.1 Features of evaluation

Feature	*Definition*	*Application*
Efficacy	Efficacy refers to the ability of a program to achieve the overall planned purpose (Gertler, Martinez, Premand, Rawlings, & Vermeersch, 2016).	Efficacy would measure whether the overall goal to reduce the use of personal cars is met.
Effectiveness	Effectiveness examines how the observed effect relates to the desired outcome by comparing the observed outcome to a baseline measure. Thus, effectiveness indicates the extent to which the intervention made a difference toward that goal (O'Donnell, 2008; Gertler et al., 2016).	Effectiveness measures the extent to which the distribution of free public transport tickets truly led to a decrease in the use of personal cars.
Efficiency	Efficiency describes the relationship between the used resources (e.g. funds, expertise; Menon et al., 2009) and the achieved success. This could require the estimation of optimization and maximization points, which reflect the use of the smallest amount of resources to gain the greatest possible level of desired output.	Assessment of efficiency focuses on the extent to which outcomes were maximized with the fewest unwanted effects in the shortest time and smallest resource use. Most efficiency measures are likely to overlap with cost-effectiveness, though would offer more scalar understanding rather than static calculations.
Fidelity	Fidelity denotes whether an intervention was implemented the way it was planned (Bradshaw & Klein, 2007; O'Donnell, 2008). A high-fidelity program would have been implemented exactly as planned, whereas a low-fidelity program would have been carried out with considerable difference from how implementation was envisioned (Bradshaw & Klein, 2007).	Fidelity would indicate whether the public transport tickets were distributed as planned.
Adaption	While fidelity is an important feature of policy interventions, in some cases adapting the original intervention may be necessary. Policies may need to be adapted for new settings, for new populations, or for changes in contextual factors, such as environmental, political, sociocultural, or economic ones (Allen, Shelton, Emmons, & Linnan, 2018). Adaption may also be required in order to promote the sustainability of an intervention (Shelton, Cooper, & Stirman, 2018).	The intervention might need to be adapted if it turns out that distribution of free public transport tickets leads to overcrowded buses during peak hours. A possible adaption could be to also encourage the use of bikes in the city by offering free minutes for bike-sharing platforms.

Feature	*Definition*	*Application*
Sustainability	Sustainability indicates the estimated life of the observed effect and whether it is expected to continue after the implementation is completed. The assessment evaluates the ability of the population addressed to maintain and manage the obtained results in the future (Fertman & Allensworth, 2016; Pollitt, 2013). Additionally, sustainability may also involve the continued assessment of the ongoing delivery or implementation of the policy over time (Shelton et al., 2018).	Sustainability would be met if the effect of the intervention (reduced use of personal cars) is maintained in the long term.
Termination	Termination refers to the reality that not all policies should be, can be, or are meant to be sustained. Over time, policies are often adapted, defunded, replaced, or fully terminated. Termination may be required if the evaluation reveals that the unintended consequences negate the intended impacts of the policy or if the policy intervention is actually causing harm.	An appropriate evaluation would also consider what impacts termination may have, such as an unwanted increase in individual use of personal cars or the reduced use of public transport requiring changes to standard schedules.

lack of sustainability soon materialized in an observed low retention rate for the placements. One in four who took a position was back to receiving cash assistance within 12 months (Glen, 2017). On a superficial and short-term level, the program achieved success in connecting individuals to jobs. However, it suffered long-term shortcomings in failing to keep people employed or placing them in jobs that could not meet their financial needs. This example illustrates that in some cases a meaningful evaluation requires assessment of both the immediate and long-term effects of a policy. Policy evaluation should not stop when the policy implementation phase comes to an end. On the contrary, *ex post* assessment should be performed to assess the sustainability of the policy (ideally at least a year after policy implementation) by using the policy platform concept described in Chapter 10.

Use of policy models to form *ex post* evaluations

As described in full detail in Chapter 10, a pragmatic model for evidence-driven policies should include a robust framework for evaluation. While it is not exhaustive, that framework covers the primary features and some additional indicators that are common in the evaluation literature (see Table 11.2). It is recommended that the parameters of the target indicators be defined in as much detail as possible during the operationalization of the issue. Definition of the parameters is crucial for ensuring that the primary objectives of the policy are tangible.

TABLE 11.2 Primary indicators for evaluation analysis

Term	*Questions*
Context[1]	What is the context in which the policy was initiated? What are the key features (size, scope, nature, behavioral barriers) of the problem or opportunity involved? Did the context change during the policy implementation? What is the final status of the context?
Purpose	What is the purpose of the policy? Are fidelity and efficacy ensured?
Timeline	What are the key dates from identifying the issues through the evaluation? How long did the implementation process last? How much time was needed before the data and the insights were available?
Resources	What resources (e.g. financial costs, human costs, social capital) were required for the policy?
Impact	What are the specific benefits or harms for individuals, groups, or populations that the policy delivered (usually compared to a baseline or status quo method)? Were the benefits and impacts equitably distributed?[2]
Reach	What proportion of the population is reached by the intervention? How representative of the overall population are the individuals participating in the intervention? What is the absolute number of people affected by the intervention?
Cost-effectiveness	What is the financial value based on what was invested in relation to total gain?
Population gains	Does the entire population benefit even when they are not targeted, affected, nor participating?
Future uses	What should stay in place? How can others use it? Will the policy continue to make an impact? Should the policy be put back into the toolkit until it becomes necessary again?

Notes:

1 See also: Allen, P., Pilar, M., Walsh-Bailey, C., Hooley, C., Mazzucca, S., Lewis, C. C., . . . Brownson, R. C. (2020). Quantitative measures of health policy implementation determinants and outcomes: A systematic review. *Implementation Science*, *15*(1), 1–17.

2 See also: Emmons, K. M., & Chambers, D. A. (2021). Policy implementation science – An unexplored strategy to address social determinants of health. *Ethnicity & Disease*, *31*(1), 133–138.

Methods and data collection

In order to evaluate the effect of an intervention, relevant data should be collected and analyzed. The main objective guiding the choice of methodology would be to adequately address the evaluation questions, while performing a fair and unbiased assessment (UNEG, 2016). The specifics of the research design would often depend on the effect that one is trying to assess and the available resources (e.g. time, evaluators). Before time and resources are invested in a large-scale evaluation, pilot studies are highly recommended. Pilot studies are usually quick and small in scale. They allow researchers to obtain an idea about the expected outcome as well as to forecast any challenges that may arise in the evaluation process (Van Teijlingen & Hundley, 2001). Equally important is to be selective and targeted. Not

everything can be evaluated, and the depth and scope of evaluations should be proportional to the size, scope, and impact of the policy (OECD, 2015a; OECD/Korea Development Institute, 2017). In addition, experiments can be an effective way of evaluating the effectiveness of policy implementation from a user perspective and can therefore be performed in a lab or a field environment.

There are numerous research designs to evaluate the success of a given intervention. Most frequently researchers conduct **experimental designs**, **nonexperimental designs**, and **economic evaluations** (see Table 11.3). The various approaches are not mutually exclusive and evaluators often triangulate different methods (e.g. quantitative and qualitative methods) to enhance reliability and ensure valid results (Menon et al., 2009).

TABLE 11.3 Research designs

Type of design	*Description and key features*	*Further reading*
Experimental design	• Manipulation of at least one predictor variable • Assess the effect of the manipulated variable(s) on (an)other variable(s) of interest • Aims to detect a cause-and-effect relationship	WHO (2007)
Randomized controlled trial (RCT)	• Considered as the gold standard methodology for evaluating an intervention • Comparison of at least two groups • Individuals are randomly assigned to an intervention or control group	WHO (2007) Behavioural Insights Team (2014) Jamieson and Giraldez (2017) Cartwright (2007)
Quasi-experimental design	• In some cases, more ethical, feasible, and cost-effective than RCTs • Compares (natural) groups without a randomized allocation	WHO (2007)
Natural experiment	• Studies policy reform as an experiment itself • Often classified together with quasi-experimental design	Blundell and Costa Dias (2002)
Nonexperimental design	• No variable manipulation • Examines naturally occurring relationships between variables (e.g. through surveys or focus groups) • Relies on observation, interpretation, or interactions • Only correlational statements are possible • Usually high in external validity	Glasziou (2004)
Economic evaluation	• Considers the cost-effectiveness and sustainability of policies • The type depends on the objective and the features of the intervention evaluated	Menon et al. (2009) Drummond, Sculpher, Claxton, Stoddart, and Torrance (2015)
Cost-minimization analysis	• Compares the costs of policies • Aims to identify the least expensive policy among two or more options that have identical benefits	Menon et al. (2009)
Cost-effectiveness analysis	• Compares the relative values of competing interventions • Costs are measured in monetary terms, and effectiveness is assessed independently	Menon et al. (2009)

(*Continued*)

TABLE 11.3 (Continued)

Type of design	*Description and key features*	*Further reading*
Cost-utility analysis	• A particular form of cost-effectiveness analysis, which is used when effectiveness or utility is hard to quantify in monetary or metric terms • Compares the consequences of policies on the basis of set criteria • It is used to compare interventions with different (nonmonetary) benefits	Dernovsek, Prevolnik-Rupel, and Tavcar (2007) WHO (2009)
Cost-benefit analysis	• Estimates the net monetary cost of achieving a particular outcome • Based on the principle that the monetary benefits of an intervention should exceed the costs of its implementation • Considers costs and benefits in monetary terms	Treasury (2013)

Common data collection methods

Along with selecting the research design, researchers make a choice about which specific data collection methods to use. The choice of which methods to use is driven by several factors: the objectives, the design of the research or the intervention, practical factors related to time and the available resources, and the type of analysis that would best capture the effect of interest. The data obtained can either be quantitative, qualitative, or mixed methods, which integrates both quantitative and qualitative methods (WHO, 2007; Palinkas & Rhoades Cooper, 2017). Quantitative data encompass information expressed in numbers (typically from surveys) and are analyzed with statistical techniques. In contrast, qualitative data are typically nonnumeric (involving text and words) and are analyzed through a separate set of techniques depending on the research question (Bryman & Burgess, 1994). In some cases, researchers may gather and triangulate both quantitative and qualitative data to test the effects of an intervention (Johnson & Onwuegbuzie, 2004). The required data may be readily available (e.g. existing texts or survey data) and may merely require collation and analysis. This is, for example, the case when using archival records or existing reports and documents of previous initiatives. More frequently, however, the data necessary to answer the research question need to be collected using primary data collection (Menon et al., 2009). Data collection initiatives typically reply on standardized instruments and questionnaires, interviews, and observations (Menon et al., 2009); ideally, the quantitative measures that are used have been psychometrically tested and validated. In recent years, Big Data approaches have become increasingly prominent (Jin, Wah, Cheng, & Wang, 2015).

Standardized instruments and questionnaires provide a common approach to obtaining information on a wide range of topics from a large number of diverse individuals. They are typically administered online and have the advantage of being relatively quick and inexpensive, which is particularly convenient when data are collected from large samples of respondents. Large samples are typically required to ensure that a given investigation has sufficient **statistical power** to detect the effect of interest. The statistical power refers to the probability that, if it is false, the null hypothesis, which predicts no significant statistical difference between the observed variables, will be rejected. Typically, the larger the effect is

and the larger the sample size is, the greater power the study has to detect a significant and real impact (Cohen, 1992). Statistical power and sample size are important considerations because inadequately powered studies can lead to false rejection of the null hypothesis or failure to detect a real effect. Equivalently, cluster randomized trials, which may be used for policy-related studies, require large numbers of clusters or units to power the study.

BOX 11.2 PRIMARY AND SECONDARY DATA

Primary data are firsthand data gathered by the researchers themselves for the specific purpose of the study.

Secondary data are data previously collected and readily available, like archival records or existing reports and documents of previous initiatives.

While standardized tools and questionnaires allow for large samples of respondents, they often provide just a general snapshot of the issue and can be lacking in the depth of information provided. In order to obtain more detailed information about individual impressions and experiences, researchers often conduct individual interviews (Menon et al., 2009). Qualitative interviews can often be used to complement and provide richer context to responses to questionnaires and surveys; interviews are also useful to provide insight into issues or phenomena that are not well understood. **Focus groups** in particular can be a quick and useful way to explore both similar and divergent points of view across diverse stakeholders (Menon et al., 2009). Nonetheless, interviews often require trained facilitators, and data collection and data analysis can be time-consuming and resource intensive.

One relative weakness of interviews and questionnaires is their frequent reliance on self-report of one's past behavior or perceptions of the environment or social context, which is known to be prone to several biases. For example, self-reports are particularly vulnerable to **social desirability bias**, which refers to a tendency of respondents to answer questions in a way they think may appear more favorably to the interviewer (Phillips & Clancy, 1972). This response bias can distort the interpretation of mean tendencies and individual differences, as it can introduce an overestimation of positive and an underestimation of negative attitudes or behaviors. Further, retrospective assessments can be prone to **cognitive biases** of recall, where memory of individual behaviors can be shaped by current moment and mood (Schacter, 2012). People are often found to misremember and provide inaccurate judgments of their performed behavior (Behavioural Insights Team, 2014). Similarly, individuals can also be poor at predicting their behavior. For instance, respondents have been found to perform much less exercise than they predicted they would (Behavioural Insights Team, 2014).

On-site observations provide a potential means to overcome some of the biases associated with self-reports. In this approach, direct observation protocols are used to evaluate how a program operates, encompassing the ongoing processes and activities, as well as the results that are observed along the course of the initiative (Menon et al., 2009). While observations allow real-time tracking of the program's implementation and progress as they occur, including the extent to which a full program is being implemented with fidelity or the extent to which it is adapted, they can be very costly and time-consuming. In order to ensure comparability of results across sites, data collectors must be trained to use the protocols in the same manner, and clear guidelines should be set to facilitate consistent interpretation of protocols from different sites.

Alternative methods to overcome these biases and test the effectiveness and impact of policies can include experiments in a controlled environment (OECD, 2017a, 2017b). Because extraneous variables can be controlled in experimental studies, and experimental research designs may be replicated, this form of research should be considered when it is likely that results from qualitative research, such as focus groups or individual interviews, are skewed due to biases or other interfering factors.

Big Data

Another potential tool for overcoming issues related to self-report that has generated substantial excitement in recent years is the use of **Big Data**: collated sets of digital footprints acquired at large volume, velocity, and variety, often matched from multiple sources (Jin et al., 2015). One of the most significant changes that the digital era has brought to policymaking is the availability of constant, user-generated streams of digital information. Digital traces such as social media posts, Google searches, financial transactions, and bus card swipes are automatically collected by various devices and constitute large, often inexpensive data sets with ecologically valid information about individual choices and behaviors. The information they contain is exceptionally rich, encompassing geographical locations, social connections, financial choices, physical activity, audio, video, etc. (Kosinski, Wang, Lakkaraju, & Leskovec, 2016). These Big Data samples can offer time- and resource-efficient opportunities to explore natural human behavior (Kosinski et al., 2016).

Big Data provide a direct means to address concerns such as those expressed by Lindblom (1959) over the limited utility of insights based on retrospective and unrepresentative evidence. Big Data sources can help to overcome such concerns by allowing regular, flexible, and granular access to larger populations (Back et al., 2010). They bring many advantages, including the opportunity to capture even small effects (given the large samples sizes often available), as well as the opportunity to obtain behavioral insights free from the potential social desirability bias associated with self-reported surveys (Kosinski et al., 2016; Matz, Gladstone, & Stillwell, 2017a). Nowadays, online patterns of behavior such as Facebook likes can be used to reliably infer personal characteristics such as ethnicity, gender, sexual orientation, political affiliation, and personality traits (Glenn & Monteith, 2014; Bachrach, Kosinski, Graepel, Kohli, & Stillwell, 2012; Lambiotte & Kosinski, 2014). The analysis of these data sets can help governments and organizations identify individuals and groups of interest, as well as population trends and risk factors. For example, analysis of the volume of Google searches for illegal substances can provide insights about interest in these substances and their popularity (Deluca et al., 2012). Such insights allow governments to monitor and predict potential public threats and assess the effectiveness of interventions designed to tackle them. Additionally, access to Big Data when implementing policies affords researchers the possibility to quickly evaluate and flexibly revise interventions according to how they are received, ultimately allowing for improved regulation (Schintler & Kulkarni, 2014). Of course, this will only *occur if the information is used* to inform the design of new policies and regulations or to modify existing ones.

The use of Big Data analytics affords numerous advantages, but governments and organizations must also address various challenges, including issues of representation, accuracy, access, and privacy. Big Data are not always accurate and balanced representations of entire populations, and failure to appreciate this can leave policy-relevant groups ignored (Ruggeri et al., 2017; Bentley, O'Brien, & Brock, 2014; Taylor & Schroeder, 2015). Big Data are not always equally accessible to all parties and often either are generated outside public administrations or are not available to all departments within administrations. Further, there are various ethical concerns about organizations and governments collecting and using Big Data to target consumer behaviors without the explicit consent of the users. Without clear ethical standards about how Big Data approaches should be implemented, there is a risk that they may be used to manipulate or disproportionately benefit specific groups (Ruggeri et al., 2017), with lower income and disparity populations at greater risk for being less likely to receive benefits. To ensure that such approaches have society's best interest in mind is a challenging but necessary task that requires clear guidelines around individual control of shared data, confidentiality, and transparency about the ways in which these data are used and accessed

(Ruggeri et al., 2017). Ultimately, one of the most important roles of Big Data in policymaking may be to demonstrate empirically that interventions capitalizing on such data result in broad public benefit (Ruggeri et al., 2017). In this way, it may prove to be a powerful tool both for generating evidence and for establishing standards for scientific insights to be used in policy.

Standards for evidence

To inspect the evidence, the evaluation design should employ a research method and data collection strategy that are rigorous and well-suited to the specific evaluation questions. This is critical because the final decision from an evaluation will be to determine whether a policy has been a good one, in some form. It is therefore important to also have standards for the valuation of evidence available in advance.

To assess the extent to which evidence is available for a topic (e.g. from a scientific study or a policy), the Cambridge Policy Research Group produced the **Index for Evidence in Policy (INDEP)** (see Figure 11.2; Policy Research Group, 2016). INDEP

FIGURE 11.2 Index for Evidence in Policy (INDEP)

assigns the evidence of interest a rating ranging from 0 (theory proposed) to 10 (impact validated), reflecting the quality, amount, and consensus regarding existing scientific evidence. While evidence of any rating could be used to inform policies, the insights with lower ratings should be treated with caution and may require additional scientific grounding. Further, INDEP also considers the evidence around the generalizability of the given insight (e.g. the settings, contexts, populations, and conditions in which the policy may be effective). As reflected in the INDEP, the evidence for the policies should not be produced only in controlled or niche environments, but ideally experiments in real-life, less-controlled, and lower resource settings should occur after an intervention is established in a controlled trial.

Scoring policies

In the United States, one of the tensest periods (for direct stakeholders, at least) is when the Congressional Budget Office (CBO) releases its estimates of financial and human impact after legislation has been presented by Congress. While the CBO is technically nonpartisan and offers only estimates, its projections often set the tone for public and political discourse about a given bill. While it is rare that public attention returns to these estimates in follow-up laws that are eventually passed, government and political structures often rely on these evaluations in both future development of prospective legislation and related debates. Most countries have similar structures in place for such legislation, but policy scoring is considerably less institutionalized, with few parallels to draw from across countries.

Much like with policy cycles, there are a substantial number of theoretical approaches to the evaluation of policy (Trochim, 2009). Most, if not all, of these frameworks present very useful information for classifying and organizing critical features, but they provide little practical direction (Howlett et al., 2017). Furthermore, many of these frameworks offer concepts for measurement without producing actual metrics or scales, nor detailing how to weigh various aspects on the basis of their overall impact on or value to outcomes. It is likely that many policymakers, researchers, and stakeholders from invested organizations will refer to these only to find general agreement about the importance of theory but will be disappointed with the lack of detail on specific application. While this is a genuine challenge, it is largely viewed as a matter of broad categorization intended for simplifying teaching and research (Howlett et al., 2017). This is not meant as a criticism of published models; it merely represents an important gap in the field, which is a clear opportunity for scientific contribution to policy.

How to understand policy evaluations

The work of Cheung and colleagues (2010) offers an exceptional glimpse into policy evaluation through a framework for scoring *policy reports*, which include the evaluations but go beyond simply the policies and tools themselves. Through eight general criteria (outlined next), they propose measures for assessing the information and approach in *health* policy reports. The criteria encompass the content, how robustly the information is provided, and which fundamental elements are included. These domains are useful not only for shedding

light on what is important in reporting but also for *a priori* thinking about and planning for policy implementation.

1 Accessibility
2 Policy background (i.e. the source of the health policy)
3 Goals
4 Resources
5 Monitoring and evaluation
6 Political opportunities
7 Public opportunities
8 Obligations

There are many examples of how to generally weight and standardize indicators relevant to policies. The OECD has been a leader in producing systematic approaches to policy evaluation through combining and standardizing social and economic indicators (see Nardo et al., 2005). However, in most cases, policy indices are established discretely between domains, though they may have significant overlap with a variety of sectors. An example of this is the Small and Medium-Sized Enterprises (SME) Policy Index, which is built for the scoring of frameworks and capacities of governments to optimize growth through supporting local business development. It is an excellent tool for a general review of SME policies. Such instruments may be relevant for cross-country comparisons in certain policy areas, but they are not a tool for robust policy evaluation.

Similarly, one of the most powerful indicators of national economic stability is inequality (Piketty & Saez, 2014; Piketty, 2000). Within this area of work, the **Gini coefficient** (a measure where 0 equals perfectly equal incomes and 1 equals perfectly unequal incomes across a country) is commonly used as a score for assessing national economic inequality, which is useful for predicting a number of likely barriers to growth and stability (Gastwirth, 1972). The Gini coefficient is widely reported in academic, government, industry, and third-sector initiatives, which is likely due to its simplicity in scoring and use. Gini scores have catalyzed a substantial number of policies aimed at reducing inequality to spur growth, but the coefficient is entirely focused on incomes and is less useful for broad application.

Drawing from an entirely different source, one measure widely reported in the media is the World Press Freedom Index, assigned annually by the organization Reporters Without Borders. This multidimensional index is able to score and standardize a number of indicators (e.g. media independence from government, legislative protections, censorship, violence toward journalists) critical to members of the press. This score is useful for understanding the media freedom within a country. Further, for countries that value freedom of the press, it can indicate potential areas for improvement.

In spite of the value that many existing scoring approaches provide in specific contexts, at present there is no scientific, systemic approach to standardized scoring of policies that can be applied across different domains. At the same time, a substantial amount of research on policy – theoretical and applied – has converged on a set of common indicators deemed critical (Allcot & Mullainathan, 2010). This presents a tremendous opportunity to standardize the policy evaluation process to help policymakers to determine the effectiveness of interventions that have already been applied or are currently being considered.

While there is no standardized scoring tool available for all policies, it is possible to utilize the extant indicators on which policy researchers have converged. It is recommended

that a common scale is used as often as possible, followed by appropriate waging similar to the approach used by the OECD. This also serves as a guide for the minimum information that should be included in high-quality policy reports, similar to the framework provided by Cheung and colleagues (2010). Ideally, such an approach would maximize the accessibility of the policy evaluation to policymakers, experts, and stakeholders, as well as to the general public.

In the example presented in Box 11.3, 20 indicators are assessed: most are scored from 0 to 5, the evidence assessment scoring ranges between 0 and 10, and some items range from negative to positive. A separate scale for evidence assessment is introduced as a means of correcting for policies where most projections are not based on empirical evidence. The purpose behind this is to provide a scale ranging from 0 to 100 that is easily understood and requires no advanced knowledge of statistics or policy evaluation. In Box 11.3, each indicator in the scoring is itemized, with a suggested framing for each score. Indicators cover populations involved, clarity of important indicators, cost and resources, critical social factors, infrastructure, and scientific quality. These indicators closely correspond to the principles employed by major international organizations involved in behavioral policy (OECD/Korea Development Institute, 2017). This approach allows for a policy to be scored *ex ante* to assess its overall impact as well as likely strengths and areas of concern across specified dimensions. For example, in the generic version in Box 11.3, items 9 to 12 could be rephrased about realistic potential as opposed to empirical outcomes. This would assist with identifying potential weaknesses of or gaps in a given policy that can then be addressed through modifications or the introduction of additional policy tools.

BOX 11.3 A GENERIC POLICY SCORING SYSTEM

Populations involved

1 Which of the following population strategies will be directly or indirectly influenced by this policy? (0 – Not at all; 1 – Indirectly; 2 – Directly; 3 – Exclusively)

• Severely impaired or disadvantaged	0	1	2	3
• Impaired or disadvantaged	0	1	2	3
• Prevent	0	1	2	3
• Sustain	0	1	2	3
• Promote	0	1	2	3

2 At what level are effects expected? (0 – Not at all; 1 – Indirectly; 2 – Directly; 3 – Exclusively)

• Rare or isolated	0	1	2	3
• Small group or tribal	0	1	2	3
• Community	0	1	2	3
• Large region or national	0	1	2	3
• International	0	1	2	3

3 High-risk or population-level approach? (0 – Not at all; 1 – Indirectly; 2 – Directly; 3 – Exclusively)

• High risk	0	1	2	3
• Population	0	1	2	3

Indicator clarity

4 Are the indicators targeted by the intervention clear? (0 – Not at all; 2 – Entirely clear)
5 Does/did the evaluation refer to the intended outcome and was there a clear reference comparison (e.g. baseline, control) for determining effect? (0 – Not at all; 2 – Entirely clear)
6 Is information about the policy accessible to the public? (0 – Not at all; 5 – Entirely transparent)

Impact and resources

7 How cost-effective is the intervention?

• Costs	Lower (2)	Same (1)	Higher (–l2)
• Effectiveness	Lower (–3)	Same (–1)	Higher (3)

8 How long will it take to go from implementation to impact? (0 – Unknown; 1 – Lag after implementation; 2 – During implementation; 3 – Lag after launch; 4 – With launch)
9 How long will the impact last or how soon will outcomes regress to mean? (0 – Additional interventions will be required immediately; 5 – Once implemented, effects should be sustained for the foreseeable future)
10 To what extent was the policy implemented as intended? (0 – Not at all; 3 – Precisely as designed)
11 To what extent did the policy achieve its intended, primary aims? (0 – Not at all; 5 – Completely)
12 To what extent was there a return on the investment? (–5 – Loss; 0 – No return; 5 – Measurable return greater than amount invested)
13 Are/were there significant risks associated with this policy? (–5 – Risks for the whole population; –4 – Risks for the most vulnerable; –3 – Significant risks to a large group; –2 – Moderate risks to a large group; –1 – Moderate risks within reason; 0 – No known risks)
14 Are there any significant trade-offs? (–5 – Significantly more harms than benefits; 5 – Only benefits, no harms)

Social considerations

15 To what extent is/was this supported by the public? (–5 – Extremely unpopular; 0 – No support or dissent; 5 – Extremely popular)
16 To what extent is the policy politicized? (5 – Completely apolitical; 0 – Explicitly biased for or against a political group)

Infrastructure

17 To what extent are there legal or regulatory structures to support this approach? (0 – No regulatory backing; 5 – Well defined with oversight)
18 Is it possible to replicate the policy in other locations? (0 – No; 5 – Directly and without modification)

Scientific quality

19 Evidence assessment: using the 0–10 scale from the PAI, rate the level of evidence in support of this intervention.

Well-being

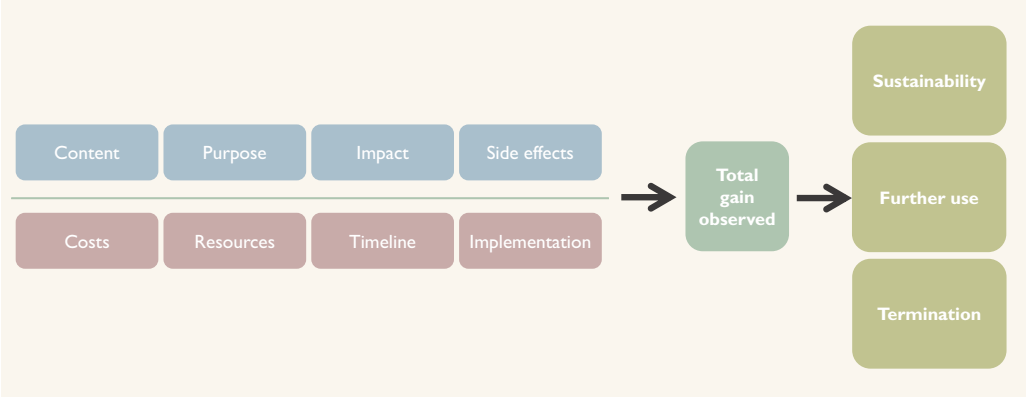
20 What is the impact on well-being? (–1 to 1 for every dimension measured)

It is important to utilize any such tool with some caution, as it is merely for the purpose of standardizing and informing policy discussions and for comparing policy options, but should not be perceived as the absolute word on policy decisions. Additionally, as earlier chapters have presented, there are strong reasons for expanding the number of critical outcomes measured. Along with well-being, these could encompass the reduction of inequalities or the increase of economic stability and physical security.

What is a good policy?

The ultimate test to deem a policy *good* reflects its success at improving desired outcomes with little or no harm. Notably, there are different paths of varying levels of efficiency and sustainability through which a policy can arrive at this effect. In the spirit of the famous quote, "If you treasure it, measure it" by Lord Gus O'Donnell (Copps, 2011), we can estimate the valuation of a policy by considering its performance across various dimensions such as context, purpose, impact, side effects, costs, resources, timeline and implementation, population gains, and future uses. Following this multidimensional approach, a good policy considers the relationship between the outcome and the necessary resources to increase the overall gain, while an even better policy has high external validity and can be translated into other domains and used in the future (Figure 11.3). Crucially, future applications may carry novel challenges, and these should be carefully evaluated before future uses are pursued and implemented (Bloom, Genakos, Martin, & Sadun, 2010). Sufficient evaluation of a policy and the consideration of its features are recommended to assess the evidence, which facilitates the statement of whether a policy is *good*. In this way, it is more likely that we can understand whether a policy has produced the most positive outcome for the greatest number of relevant individuals, groups, and populations.

FIGURE 11.3 Valuation of a policy



Essay questions

1 If you were only given 5 minutes to present the implementation plan for a major policy, which elements would you focus on and why? Give a policy example to illustrate.

2 Think of a possible example for a policy and explain the features of evaluation based on this example. (Hint: Use the examples from Chapters 4–8.)

3 Imagine that you are part of an evaluation group and you want to support individuals to stop smoking. Name a possible policy and decide which research method you would use to assess the success of the implementation. Explain.

4 A previously implemented policy was terminated. You are asked to rate the policy. How would you approach this and on which points would you base your rating? Explain.

5 What is the benefit of failed policies or having a very bad approach to a major challenge? How can evaluation help?

6 What are the implications of not planning an evaluation early in the process of developing a policy?

7 Describe five mistakes that could be made in designing a policy evaluation.

8 How might policy evaluations differ for various domains, such as health policy, energy policy, and school policy?

9 If a policy evaluation was interested in assessing whether the policy and its impacts were equitably distributed, how might you go about determining this?