What Questions Do Evaluations Answer?

B. Essama-Nssah

Evaluation can be seen as an inferential process designed to produce evidence to answer a set of programmatic or policy questions. An evaluation question is a request for information by decision makers and other stakeholders about the performance or results of an intervention. Well-posed questions are a prerequisite for useful evaluation because they focus the evaluation on issues that stakeholders care about, and determine the appropriate evaluation design and methods. All evaluation questions can be answered on the basis of three types of inferences: descriptive, associational, and causal. This note frames evaluation within the logic of asking and answering questions. It provides guidance on asking the right questions, and considers the basis of credible answers. The credibility of an evaluation hinges critically on the validity of the underlying inferences. The validity of these inferences depends on the rigor of the underlying logic, and the reliability and validity of the supporting measurement system.

Over the past three decades or so, pressure has been mounting on governments around the world to improve the effectiveness of public policy. This has led to the emergence of a public sector management approach known as Managing for Results, which has propelled evaluation to the forefront of evidence-based decision making. Effective evaluation can produce reliable information on what works, what does not, and why. Policy makers may use this information to modify or cancel ineffective programs, or design new ones and thus make the most of limited resources (Grossman 1994). Within this informational perspective, evaluation can be viewed as an inferential process designed to produce credible evidence to answer a set of well-posed policy questions (Gertler et al. 2011). Indeed, Ravallion (2009) notes that: “The art of good evaluation is to ask the right questions at the outset, motivated by existing knowledge gaps and to tailor the data and analysis to answering those questions in the specific context.” This vision creates the need to know: (i) how to ask the right questions, (ii) how to ask questions right, and (iii) how to find the right answers to the questions asked.

This note frames policy evaluation within the logic of interrogation and provides guidance on asking and answering questions in areas of concern to decision makers and other key stakeholders. Evaluation questions provide an essential mechanism by which stakeholders request information about performance or results. The note proposes template questions that can guide the formulation of specific questions. These questions presume that policy makers would like to know: (i) whether they are doing the right things the right way, (ii) if what they are doing is working and worth the cost, and (iii) what explains the observed outcomes.

An evaluation question is answerable only if there is a logical and feasible pathway to an answer. All evaluation questions are answerable on the basis of three types of inferences: descriptive, associational (or relational), and causal. The credibility (and hence the usefulness) of an evaluation hinges critically on the validity of these underlying inferences, which in turn, depend on the rigor of the underlying logic, and the validity and reliability of the supporting measurement system.
Asking the Right Questions

Are we doing the right things?
This question focuses on the rationale of the program. A program rationale answers the question: Why this intervention? A positive answer to this question is: Because it is the right thing to do for the target population, given the problem and the circumstances they face. Maintaining and improving the living standard of the population is the ultimate goal of public policy and a fundamental expectation of the governed (Sen 1987). In this broader context, programs are the mechanisms through which we take public policy to the target population. The question of whether we are doing the right things concerns both the objectives and the way we pursue those objectives. It raises the issue of the soundness of the logical connections between program activities and the intended results. The quality of this logic depends on the quality of available knowledge about the best way to achieve the intended outcomes. This knowledge helps determine whether the program is appropriate or plausible. Therefore, saying that a given program is “the right thing to do” means two things: (i) the program as a whole is aligned with the overall policy goal, and (ii) the selected activities are among the best possible means to achieve the desired ends.

The identification of goals, target groups, and strategies emerges from a needs assessment that explores the nature and extent of the problem to be addressed, including its root causes and social consequences. It also identifies the needs and characteristics of the target population (for example, area of residence and other sociodemographic characteristics). This information can help calibrate the intervention in terms of types of services needed, intensity, and duration. Furthermore, a needs assessment provides baseline information that can be used to judge program performance and progress toward goals. It can also be part of an evaluation of an ongoing program to determine whether the program remains relevant to current conditions.

The following are illustrative questions that arise in the context of a needs assessment:
(i) What is the nature and the extent of the problem to be addressed?
(ii) What are the characteristics of the affected population?
(iii) What services are needed?
(iv) What is currently being done to address this problem?
(v) What service delivery mechanisms are most suitable for providing services to the target population?

A needs assessment is analogous to a poverty assessment. Just as the latter underpins a poverty reduction strategy, so does the former a program designed to address a social problem. The design and implementation of effective safety nets, for instance, require the identification of the vulnerable, the source of vulnerability, and feasible coping mechanisms. Grosh et al. (2008) offer a framework for conducting this type of needs assessment. Gantner (2007) provides an assessment of the need for PROGRESA (Programa de Educación, Salud, y Alimentación) in Mexico. The main findings of the assessment are summarized in box 1.

Are we doing things right?
Having established that the intervention is a good idea, the next logical step is to check that it is implemented as designed. Until implementation, the relationship between an intervention’s inputs, outputs, and expected outcomes is merely hypothetical. It is assumed that a sound intervention plan is most likely to lead to desired outcomes if implemented as designed. The question of whether we are doing things the right way focuses evaluation on the performance of the intervention delivery. Provision, utilization, and coverage represent three critical dimensions of service delivery for the achievement of desired outcomes. These aspects underpin the logical connection between inputs and outcomes. Provision ensures that services are available and accessible in sufficient quantity and good enough quality to the target population. One can further characterize the adequacy of provision in terms of frequency and intensity. These issues give rise to the following evaluation questions:

- **Frequency**: Is the treatment administered often enough to make results likely? For example, the success of an agricultural extension program depends on the frequency of visits to farmers by the extension workers.
- **Intensity**: Does the level of concentration of the intervention make the results likely? For instance, a performance evaluation of the health component of PROGRESA revealed that some beneficiaries received only a fraction of the daily ration of nutritional supplements they were supposed to get from the program.

**Box 1. The Need for PROGRESA**

This program, based on an integrated approach to poverty reduction (involving education, health, and nutrition), started in 1997, focusing initially on rural poor. It was extended to the urban poor in 2002 when its name changed to Oportunidades. The emergence of PROGRESA means that social policy makers in Mexico at the time believed that they were not doing the right things in the fight against extreme poverty. It has become clear that targeted subsidy programs such as those targeting tortillas (Tortibono) and milk (Liconsa) were not helping the poorest of the poor (Lustig 2011). PROGRESA was therefore designed to deal with some of the limitations of these programs while embracing an innovative approach to human capital development aimed at reducing both present and future poverty.

Source: Author’s compilation.
The target population receiving the intervention must accept it and make use of the services provided (utilization) as prescribed for the expected results to occur. Coverage (the proportion of the target population reached by the program) is a result of the interaction between the supply and demand sides of the intervention. It is intimately linked to the issue of targeting as raised by the following question: Is the program targeted at a specific and appropriate segment of the population? These considerations imply that to enhance the chance of achieving the desired results, the intervention must be appropriate, adequate, and targeted to the right segment of the population who will make use of it as prescribed. Box 2 presents the findings of an assessment of the quality of targeting in the context of PROGRESA.

**Is the intervention working?**

Programs are designed to solve some social problem by addressing the needs of the target population using the best strategies available. If we are doing the right things the right way, the expectation is that intended results will occur. It is therefore important to clarify at the outset what success would look like and how we will know if it has been achieved or not (Mackay 2010). The question of whether the intervention is working has two basic components (McDavid and Hawthorn 2006). First, the evaluation needs to determine whether and to what extent the program achieved the observed outcomes. Second, the evaluation needs to confirm that the outcomes observed are consistent with those intended.

The question of whether a program did produce the observed outcomes raises the issue of attribution. There are several ways of framing this issue, as demonstrated by the following questions:

(i) What is the net effect of the intervention?
(ii) Did the program work?
(iii) How much of the impact is due to the intervention?
(iv) What would have happened without the intervention?

Attribution is predicated on the notion that the intervention (or cause) is both a necessary and sufficient condition for the outcome (or effect) to occur (DIFID 2012). In this perspective, the intervention is considered as a single cause associated with the observed net impact that can be attributed to it. Necessity implies that, without the intervention, the effect would not be observed, while sufficiency means that with the intervention, the outcome of interest changes more than it would have changed without it.

The fact that a development intervention is an open system that interacts with its environment or context to produce effects means that such an intervention must be viewed as part of a causal package, wherein several causes work together to produce the effect. Asking whether an intervention is working is thus equivalent to asking whether it is making a difference and contributing to bringing about the observed effects. The evaluation question can therefore be reformulated as: *Is the program making a difference?* This perspective is more relevant to the evaluation of the impact of complex interventions and requires particular focus on the role of the intervention within the package under consideration (DIFID 2012). Box 3 presents a list of some of the attribution questions considered in past evaluations of PROGRESA.

**Is the program worth the cost?**

This question focuses on the relationship between the costs of a program and its key outcomes or benefits. The question is meaningful only if it can be established that the benefits under consideration are attributable to the program under evaluation. Linking costs to benefits can be done either through cost-effectiveness analysis or through cost-benefit analysis. The key indicator in cost-effectiveness analysis is the cost-effectiveness ratio. This is the ratio of total cost to a measure of any quantifiable outcome that is important relative to the program’s objectives (Cellini and Kee 2010). In that formula, total cost is in dollars, while the outcome is expressed in its natural units, also referred to as units of effectiveness. For instance, if a program is designed to save lives, then the units of effectiveness are the number of lives saved. Thus, the cost-effectiveness ratio measures dollars per life saved. Given a desired level of outcome, the cost-effectiveness ratio is useful in determining from a set of interventions which one achieves that outcome at the lowest cost. Alternatively, one can fix the cost and use the cost-effectiveness ratio to determine which program attains the greatest outcome for the given cost.

---

**Box 2. An Evaluation of the Operational Performance of PROGRESA**

Adato, Coady, and Ruel (2000) present a comprehensive evaluation of the operational performance of PROGRESA, which is an excellent example of how to analyze performance issues in the context of a large and complex program. One of the issues considered in that evaluation is whether targeting was conducted properly. Under PROGRESA, eligible villages were remote and had to have at least one primary school and a health center and score high on a village marginality index, which was computed based on a preintervention survey conducted in 1997. Within treatment communities, household eligibility was determined on the basis of a poverty index using data on income and assets from the 1997 survey. This performance evaluation found that operational failures linked to the collection of data used to determine eligibility led to erroneous identification of eligible households. Consequently, many poor people were wrongly excluded from the program.

(Source: Author’s compilation.)
Box 3. Attribution Questions about PROGRESA

PROGRESA has served as a template for the design of conditional cash transfer (CCT) programs in many developing countries around the world, and a proving ground for a variety of methods of impact evaluation. In collaboration with the Mexican government, the International Food Policy Research Institute (IFPRI) has prepared a rigorous review of the impacts of PROGRESA on poverty, education, health and nutrition, and on households’ pattern of decision making. The following are some of the attribution questions considered in that review (IFPRI 2002):

(i) Does PROGRESA reduce levels of poverty?
(ii) To what extent does schooling increase as a result of PROGRESA?
(iii) Does PROGRESA affect drop-out rates, progression through grades, grade repetition, and school reentry rates?
(iv) Does PROGRESA affect school performance?
(v) What is PROGRESA’s impact on the trade-off between child education and labor?
(vi) What has been the impact of PROGRESA on adult and child health?
(vii) Has PROGRESA had an impact on the use of health clinics?
(viii) Are PROGRESA nutritional supplements having an impact on child growth?
(ix) Does PROGRESA have an effect on a household’s pattern of decision making or work?

Source: Author’s compilation.

In a manner analogous to cost-effectiveness analysis, cost-benefit analysis compares the total cost in dollars with the dollar value of program benefits. Thus both costs and benefits are expressed in monetary terms. A common indicator used in this context is the net benefit, which is equal to total benefits minus total cost. This measure helps determine whether and the extent to which the total benefit from a program exceeds the cost. The only costs and benefits that matter for the computation of the relevant measures are marginal or incremental costs and benefits. These are the costs and benefits that would accrue over and above those that would have accrued in the absence of the program (that is, in the counterfactual state).1

What explains the observed results?

Understanding what works under what conditions is the cornerstone of evidence-based policy making. Evaluation can thus support policy learning by producing knowledge about development effectiveness that policy makers can rely upon to improve current and future policies and programs. These considerations lead to further evaluation questions that request evidence to inform policy learning:

(i) How is the intervention making a difference?
(ii) Can the intervention work in a different setting?

Explanation requires a model that translates the understanding of the technical and behavioral relationships that determine the transformation of program inputs into outputs, and then outputs into outcomes. An intervention based on such relationships is an open system that interacts with contextual factors to produce the intended results. Contextual factors represent local socioeconomic, political, and environmental conditions affecting program performance and outcomes. To sort out the role of the intervention relative to such factors requires an analysis of the causal mechanisms involved based on solid theoretical and empirical knowledge. Within this framework, an explanation of why intended results were not achieved can be traced to either design failure (that is, the program is a bad idea) or implementation failure (for example, a good idea poorly implemented). The failure of early social programs based on general subsidies to move people out of poverty is certainly an instance of design failure, to the extent that those programs did not properly take into consideration the incentives facing the actors involved. In the case of agricultural extension services, if extension workers deliver the wrong package of services to farmers, that would be a case of implementation failure.

Asking Questions Right

If evaluation is about asking and answering policy and programmatic questions that stakeholders care about, then one must develop evaluation questions in a way that accounts for the opinions, interests, concerns, and priorities of the stakeholders. The formulation of evaluation questions includes specifying the dimensions of performance or results at issue and the standards by which to judge them. This requires a sound understanding of the program under evaluation to ensure that the evaluation addresses valid needs and the questions focus on the relevant issues regarding context, design, implementation, or outcomes (Brandon 1998). The program logic model is a very useful tool in this context: it provides a foundation for a comprehensive and meaningful evaluation (Frechtling 2007). Jointly developing or examining the logic model with stakeholders can help create a common understanding of the socioeconomic problem that the program is trying to solve and the strategies used to achieve the program’s goals. The process also helps identify critical assumptions that may then become a basis for the identification of evaluation questions. A clear understanding of the information needs of decision makers and other stakeholders can help make questions relevant and answerable.
The Basis of Credible Answers

Evaluation is supposed to inform policy making by providing credible evidence to answer questions that stakeholders care about. There are two types of credibility: substantive and presentational. Credibility of substance refers to the quality of the content of the evidence produced and that of the process of gathering that evidence (Chelimsky 1995). Substantive credibility underlies the objectivity of the evaluation, while credibility of presentation reflects clarity and precision in communicating key findings and conclusions to stakeholders. Both types of credibility are necessary for policy makers to respect an evaluation’s findings.

The ability to provide a credible answer to a question hinges critically on knowing what constitutes a valid answer to the question posed. For instance, a core question in evaluation seeks to determine whether and the extent to which what was supposed to happen did, in fact, happen. Most, if not all, of the questions discussed above do fit this mold. Answering this core question entails measuring what does happen, comparing it with what was supposed to happen, and explaining any significant differences between the two states (Frechtling 2007). As this core question reveals, an evaluation question may ask that performance or results be described, judged, or explained. Thus, all evaluation questions are answerable on the basis of three types of inferences: descriptive, associational, and causal.

The United States Government Accountability Office (GAO 2012) notes that the quality of inferences drawn in the context of an evaluation is commensurate with the power of the design and the quality of available data. Ultimately, the answer to an evaluation question depends on a set of constraints defined by time, money, expertise, and data.

Design

The purpose of a design is to structure the process of generating the evidence to answer a specific evaluation question. The logic of a design stems from the type of evidence needed to answer the question. For instance, answering a cause-and-effect question requires evidence that: (i) the cause preceded the effect, (ii) the cause is associated with the effect, and (iii) there is no other plausible explanation for the effect but the cause. To answer the question of whether an intervention caused a change in a development outcome, it must first be shown that its implementation preceded the change in outcome. Second, it must be shown that the change occurs where and when the program is implemented. However, the observed change cannot be attributed to the intervention on the basis of this association alone, since the association could be driven by other factors in the socioeconomic environment known as confounders. Therefore, to be sure that the observed change in outcome is due to the intervention, there must also be evidence that if the program is not implemented, the change does not occur. This requires an estimation of the counterfactual.

The logic of causal inference establishes a hierarchy among the three types of inferences used in evaluation (figure 1). Description entails drawing conclusions about unobserved parts of the world on the basis of what is observed or measured. Association concerns the existence and strength of the correlation between the intervention and the outcome. Causation, which is at the top of this hierarchy, attempts to rule out all plausible explanations of the association between the cause and the effect. Causal inference entails associational inference, which in turn assumes that both the cause (intervention) and effect (outcome) can be described (measured). Clearly, description plays a foundational role in all evaluations (performance or impact).

The outcome of an intervention depends on the degree of exposure and the effects of the confounders. The basic strategy for attributing a change in outcome to an intervention is to change exposure while holding all confounding factors constant. One way of implementing this strategy is to compare the outcome for program participants to that of a group of nonparticipants who are similar to participants in all respects except for exposure to the intervention. There are both experimental (for example, randomization) and nonexperimental ways of estimating the effect of an intervention.2 Forthcoming note 26 in this series (Essame-Nssah forthcoming) reviews two basic frameworks for resolving the attribution problem in the context of program evaluation.

Measurement

Evaluation design provides the logic underlying the answers to the policy or programmatic questions posed; measure-

---

**Figure 1. The Hierarchy of Inference**

Source: Author’s illustration.
...information needs of decision makers into requests for questions play a central role as a mechanism for translating problems and the circumstances they face. Evaluation is the best thing to do for the target population given the purpose is to inform policy makers as they decide on what under consideration has improved. But the immediate standard of the population targeted by the intervention purpose of evaluation is to determine whether the living conditions of the population have improved. As an integral part of public policy making, the ultimate purpose is to inform policy makers as they decide on what is the best thing to do for the target population given the problems and the circumstances they face. Evaluation questions play a central role as a mechanism for translating the information needs of decision makers into requests for information and in determining the choice of appropriate design and methods. There is a hierarchy of evaluation questions, moving from issues about program design, through implementation, to results (outcome and impact). This hierarchy suggests the need to integrate performance and impact evaluations in order to better explain results. In other words, to fully understand the observed results, one also needs to consider the logic underlying the intervention along with the adequacy of implementation. Similarly, there is a hierarchy among the types of inferences used in answering evaluation questions. This hierarchy reveals the foundational role of description, and hence measurement, in evaluation. Getting policy makers to pay attention to evaluation results requires credible evidence to support more conclusive answers to the questions they care about.

**Feasibility**

The application of sound logic to valid and reliable measurements leads to relatively unambiguous conclusions about program performance or results. However, the choice and application of design and methods take time and resources, and are subject to political influence. Bamberger, Rugh, and Mabry (2012) discuss time, budget, data, and political constraints that may affect the conduct of evaluation and propose strategies to cope with them. The key point here is that there is a trade-off between cost and rigor. The cost of answering an evaluation question may be so high relative to available resources that the question is effectively unanswerable.

**Concluding Remarks**

As an integral part of public policy making, the ultimate purpose of evaluation is to determine whether the living standard of the population targeted by the intervention under consideration has improved. But the immediate purpose is to inform policy makers as they decide on what is the best thing to do for the target population given the problems and the circumstances they face. Evaluation questions play a central role as a mechanism for translating the information needs of decision makers into requests for...
2. See Bamberger and White (2007) for a summary discussion of commonly used designs in impact evaluation and the associated implementation challenges.

References

This note series is intended to summarize good practices and key policy findings on PREM-related topics. The views expressed in the notes are those of the authors and do not necessarily reflect those of the World Bank. PREMnotes are widely distributed to Bank staff and are also available on the PREM Web site (http://www.worldbank.org/prem). If you are interested in writing a PREMnote, email your idea to Madjiguene Seck at mseck@worldbank.org. For additional copies of this PREMnote please contact the PREM Advisory Service at x87736.