Center for Global Development

BRIEFS

Good Quality Evaluations for Good Policy: Findings and Recommendations from Aid Agency Evaluations in Global Health

8/21/17 (https://www.cgdev.org/publication/good-quality-evaluations-good-policy-findings-and-recommendations-aid-agency-evaluations)

William Savedoff , Janeen Madan Keller and Julia Goldberg Raifman

Evaluations are key to learning and accountability yet their usefulness depends on the quality of their evidence and analysis. This brief summarizes the key findings of "Evaluating Evaluations," a CGD Working Paper that assessed the quality of aid agency evaluations in global health. By looking at a representative sample of evaluations—both impact and performance evaluations—from major health funders, the study authors developed 10 recommendations to improve the quality of such evaluations and, consequently, increase their usefulness.

Good quality evaluations are needed for good policy

Overseas development assistance has always been subject to evaluation but, in recent years, an international consensus has emerged over the need for more and better evaluation. Agencies have increased funding for evaluations and implemented new evaluation policies. While the number of evaluations has increased, we know relatively little about their quality and use. A few public reports—focused primarily on US government agencies—find that less than half of evaluations meet high standards of quality. Reports at the World Bank and the UK's Independent Commission on Aid Impact raise concerns that evaluation results are not widely utilized. These two issues—quality and usage—are likely to be related. A poor-quality evaluation is less convincing to policymakers, project managers, and constituents. To fill in some of these gaps, the CGD study is—to our knowledge—the first assessment of the methodological quality of global health program evaluations across different funders.

Around the world, tight and over-stretched budgets continue to put pressure on development assistance programs. Earlier this year, the US Administration proposed major cuts to foreign aid spending. Now more than ever, better evidence about what works in global development is needed to drive decisions on where to allocate resources to achieve results and value for money. While there has been tremendous progress over the past decade, it is critical to redouble efforts to ensure evaluations are not only of good quality but also relevant to policy.

What is a good evaluation? Relevance, validity, and reliability

To assess the quality of aid program evaluations, "Evaluating Evaluations" focused on three criteria, grounded in the methodological literature, that are essential to the quality of the evidence: *relevance*, *validity,* and *reliability*. The authors defined evaluations as *relevant* if the evaluation addressed questions related to the means or ends of an intervention, and used appropriate data to answer those questions. Evaluations were considered *valid* if analyses were methodologically sound and conclusions were derived logically and consistently from the findings. Evaluations were considered *reliable* if the method and analysis would be likely to yield similar conclusions if the evaluation were repeated in the same or similar context.

While acknowledging that funders conduct or commission evaluations with different aims, the study argues that these standards apply to all kinds of evaluations regardless of their questions and methods. For example, the sample included both "impact evaluations"—whose primary purpose is to measure and attribute impact to a particular intervention—and "performance evaluations"—that focus on questions related to such things as managerial efficiency, outputs, or beneficiary satisfaction.

The relevance, validity, and reliability of an evaluation affects whether its findings are helpful and should be heeded—whether the reader is an agency administrator trying to improve managerial efficiency; a parliamentarian trying to judge whether foreign aid is used appropriately; or a government official trying to design a new policy.

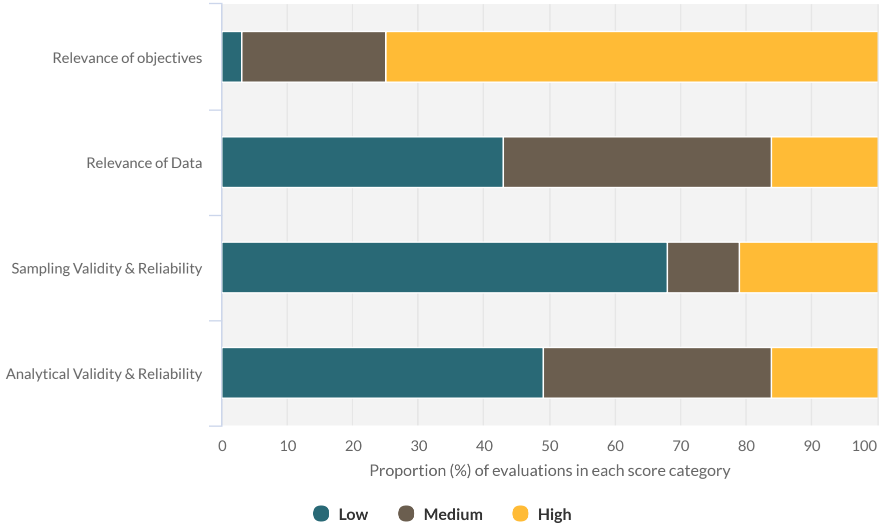
How do evaluations standup?

"Evaluating Evaluations" assessed a sample of 37 global health program evaluations. The sample was randomly drawn from a list of 299 global health program evaluations that were published between 2009 and 2014, and conducted or commissioned by five major funders of development assistance for health—United States Agency for International Development (USAID), the Global Fund to Fight AIDS, Tuberculosis, and Malaria (the Global Fund), the President's Emergency Fund for AIDS Relief (PEPFAR), the United Kingdom Department for International Development (DFID), and the International Development Association (IDA) at the World Bank. Two reviewers independently assessed each evaluation in the sample and finalized responses after discussion and reconciliation.

The sample contained many high-quality evaluations, demonstrating that good evaluations are possible using a variety of methods and in different contexts. However, the share of high-quality evaluations was also relatively small, showing considerable opportunity for improvement (Figure 1).

Summary Scores for Evaluation Quality (N=37)

Note: Percentages may add to more than 100 percent due to rounding | Source: Evaluating Evaluations, 2017



Center for Global Development

Relevance of data and objectives

Most evaluations had *relevant objectives*. Three-quarters of the evaluations addressed questions that strongly reflected the main objectives of the program being evaluated. However, evaluations were much less likely to have *relevant data*. Only 16 percent of the evaluations collected information that was fully relevant to the questions they posed. About 40 percent were judged to at least "partially" use data that addressed their main research questions. The remainder of the evaluations lacked such data.

Validity and reliability of data collection

One-fifth of the evaluations met the highest standard for *sampling validity and reliability*, 11 percent scored in the middle range, and 68 percent had a low score. To reach the highest score, evaluations had to meet several criteria including clearly justifying their data collection methods and using unbiased data sources. Most of the evaluations with the lowest score used convenience sampling without justification or lacked information on data collection methods.

Validity and reliability of analysis

Finally, we assessed *analytical validity and reliability* primarily in terms of whether evaluations addressed potential confounding factors and adequately explained their methods of analysis. Sixteen percent of the evaluations had high analytical validity and reliability. Another 35 percent scored in the middle range and 49 percent received a low score.

Baseline information, data collection, and overall analysis

Better scoring evaluations tended to have access to baseline information, used data collection methods that paid attention to bias, and conducted analyses that explicitly drew on the information collected. For example, the lack of baseline information affected an evaluation of a large-scale fund to address infectious diseases in Myanmar, without which basic trends in disease prevalence and intermediate behavioral indicators could not be tracked or analyzed. Similarly, without baseline information, evaluations that focused on project implementation or management practices had to ask key informants to recall how processes had changed rather than compare their practices to information gathered during baseline interviews.

Data collection methods also affected the strength of findings. For example, a performance evaluation of a program to improve treatment of maternal and childhood illnesses in Madagascar used a multi-stage random sample of community health workers, which provided convincing information about implementation and effectiveness of training. By contrast, a performance evaluation of an orthopaedic training program in Tanzania relied on information from respondents who volunteered to participate—a possible source of bias which in turn limited the scope of what the evaluation could conclude about implementation and effectiveness.

Finally, most evaluations relied on a mix of information sources but many analyses failed to show how they derived their conclusions from that information. For example, evaluations of a reproductive health program in India and of a water and sanitation project in

Ecuador reached conclusions that were not explicitly linked to the information sources they listed. As a result, the credibility of the findings essentially rested on the judgment of the researchers and could not be validated by retracing their analyses.

Impact and performance evaluations

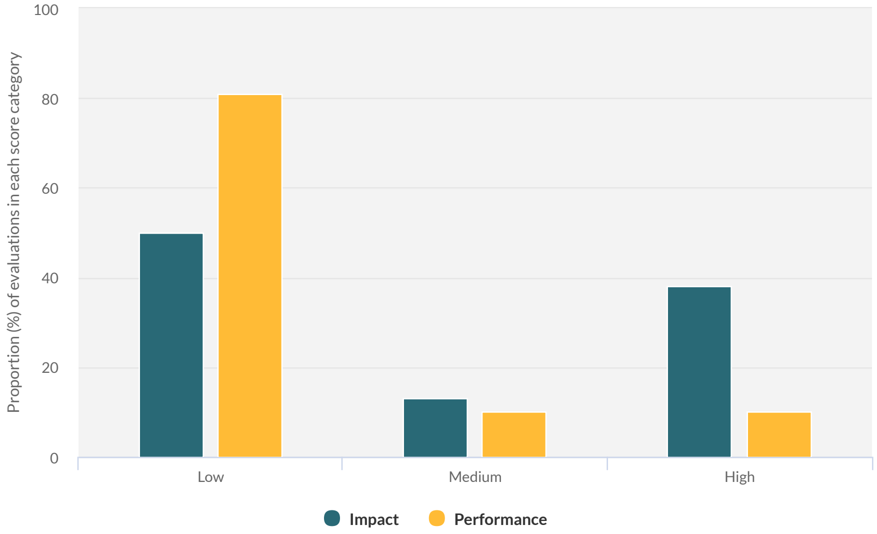
A great deal of the current debate over aid evaluations treats randomized controlled trials, or RCTs, as if they were displacing other forms of evaluation. To the contrary, less than 2 percent of the 299 evaluations identified by the study were RCTs. When considering all impact evaluations using experimental or quasi-experimental methods, the study found that these accounted for only about six percent of all evaluations; performance evaluations comprised the vast majority.

The impact evaluations generally scored better than performance evaluations on the validity and reliability criteria (Figure 2). About 38 percent of the impact evaluations and 10 percent of the performance evaluations scored high on sampling validity and reliability. Impact evaluations were also more likely to have high analytical validity and reliability—31 percent compared to 5 percent.

This difference is partially explained by the fact that most impact evaluations rely on methods that explicitly address potential bias through statistical sampling techniques. But performance evaluations, including those using qualitative methods, also have ways to collect information that enhance validity and reliability. For example, qualitative studies can address recall bias by comparing end of project perspectives with those expressed in baseline interviews.

Sampling Validity and Reliability by Evaluation Type (N=37)

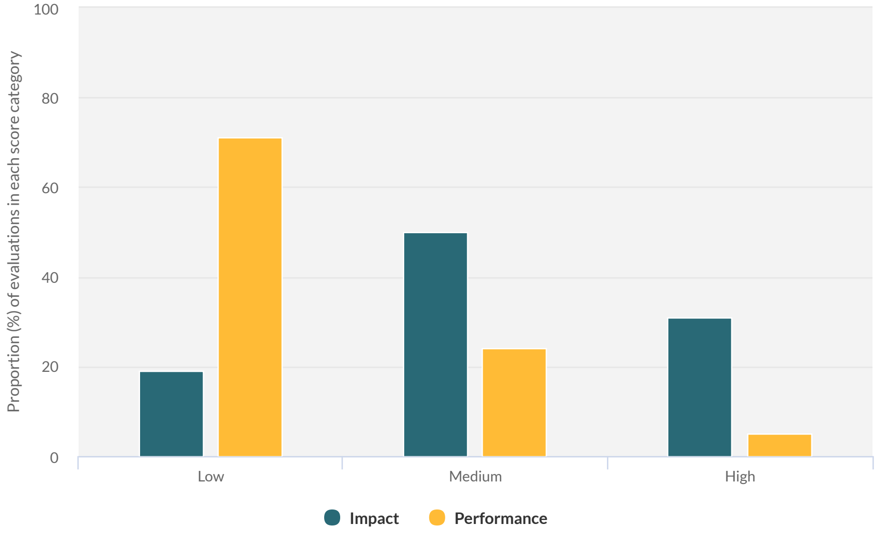
Note: Percentages may add to more than 100 percent due to rounding | Source: Evaluating Evaluations, 2017



Center for Global Development

Analytical Validity and Reliability by Evaluation Type (N=37)

Note: Percentages may add to more than 100 percent due to rounding | Source: Evaluating Evaluations, 2017



Center for Global Development

10 recommendations for improving evaluation quality

Based on this assessment, the study offers 10 recommendations for agency staff overseeing and managing evaluations to improve quality.

1. *Classify the evaluation purpose*: Be clear at meta-level data (e.g. title, abstract, tagging categories on the agency website) regarding whether an evaluation is (1) measuring the impact of a particular intervention and seeking to attribute causality or (2) asking questions about a program's performance, such as effective implementation or management.

2. *Discuss evaluator independence*: Explicitly acknowledge the evaluators' institutional affiliation and any financial conflicts of interest, along with the roles of implementers in sampling, data collection or writing reports, so that readers can assess potential bias.

3. *Disclose costs and duration of programs and evaluations*: Provide information on cost and duration for both the evaluated program and the evaluation.

4. *Plan and design the evaluation before program implementation begins*: Taking the time to develop an evaluation timeline including plans for data collection before implementation begins can improve evaluation quality (the study confirmed that this association was statistically significant) without necessarily increasing costs.

5. *State the evaluation question(s) clearly*: A clear and well-defined question helps researchers identify the right kinds of data and select an appropriate methodology; it also helps readers understand the evaluation and assess the quality of the findings.

6. *Explain the theoretical framework*: Evaluations are easier to assess if evaluators are explicit about the theoretical framework that underlies their analysis.

7. *Explain sampling and data collection methods*: Evaluations should be clear enough about sampling and data collection methods that subsequent researchers could apply them in another context, and that readers can judge the likelihood of bias.

8. *Improve data collection methods:* Interviews with key informants, focus groups, and beneficiaries are rich sources of information for evaluators but many evaluations rely on selecting informants or respondents based on convenience rather than a purposeful or random sampling approach that provides greater confidence in findings.

9. *Triangulate findings:* Development programs are by nature complex, making it unlikely that an evaluation using a single method or source of data can be complete. Combining a mix of quantitative and qualitative information can help contextualize and corroborate analyses.

10. *Be transparent on data and ethics:* Publishing data in useable formats is helpful to the evaluators conducting an evaluation, to peer reviewers judging the robustness of findings, and to readers trying to assess an evaluation's credibility. Appropriate measures should be taken to protect privacy and assure confidentiality.

Conclusion

The research shows that it is possible to assess evaluation quality across aid agencies drawing on a systematic sample and using a common set of applicable standards. Despite efforts to develop a relatively homogeneous representative sample, the range of evaluations in terms of overall aims, methods, and institutional standards remains wide making it hard to extrapolate from a relatively small sample. In addition, agencies have been implementing new evaluation policies and practices since the evaluations included in this study were published.

Despite these caveats, "Evaluating Evaluations" demonstrates that good evaluation is possible even though only a small share of evaluations in the sample met social science standards for research. Improvement is possible, often within existing budgets, through better planning and systematic attention to data collection. Improving evaluation practice to assure relevance, validity, and reliability can improve the evidence needed to ensure that development programs contribute to better lives around the world.

Further Reading

CGD Evaluation Gap Working Group. 2006. *When Will We Ever Learn? Improving Lives through Impact Evaluation*. Co-chaired by William Savedoff, Ruth Levine, and Nancy Birdsall. Washington, DC: Center for Global Development.

Government Accountability Office. 2017. "Foreign Assistance: Agencies Can Improve the Quality and Dissemination of Program Evaluations." Report to Congressional Requesters, GAO-17-316, Washington, DC: GAO. March.

Hageboeck, M., M. Frumkin, and S. Monschein. 2013. "Meta-evaluation of quality and coverage of USAID evaluations 2009-2012." Arlington, VA: Management Systems International.

Independent Commission for Aid Impact. 2014. "How DFID Learns." Report 34. London, UK: DFID.

OECD Development Assistance Committee. 2010. "Quality Standards for Development Evaluation." DAC Guidelines and Reference Series. Paris, France: OECD.

Raifman, Julia Goldberg, Felix Lam, Janeen Madan Keller, Alexander Radunsky, and William Savedoff. 2017. "Evaluating Evaluations: Assessing the Quality of Aid Agency Evaluations in Global Health." CGD Working Paper 461. Washington DC: Center for Global Development.