Published in final edited form as:

J Dev Effect. 2018 ; 10(2): 277–289. doi:10.1080/19439342.2018.1452779.

How well are aid agencies evaluating programs? An assessment of the quality of global health evaluations

Julia Raifman^{a,*}, Felix Lam^b, Janeen Madan Keller^c, Alexander Radunsky^d, and William Savedoff^{iD,c}

^aDepartment of Epidemiology, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA

^bMonitoring and Evaluation, Clinton Health Access Initiative, Boston, MA, USA

^cCenter for Global Development, Washington, DC, USA

^dDepartment of Global Health and Population, Harvard T.H. Chan School of Public Health, Boston, MA, USA

Abstract

Evaluations are key to learning and accountability. We assessed the methodological quality of 37 randomly selected programme evaluations from 5 major global health funders. Two researchers rated each evaluation for relevance, validity, and reliability and met to resolve discrepancies. Most evaluations asked questions relevant to the health programme, but less than 40 per cent of impact

CONTACT Julia Raifman, JuliaRaifman@gmail.com, Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, 615 North Wolfe Street, Room E7133, Baltimore, MD 21231, USA.

*Present affiliation for Julia Raifman is Department of Health Law, Policy & Management, Boston University School of Public Health, Boston, MA, USA.

ORCID

William Savedoff <http://orcid.org/0000-0001-5313-2013>

Supplementary material for this article can be accessed here.

Julia Raifman, ScD, did her postdoctoral research at Johns Hopkins and moved into an assistant professor position in the Department of Health Law, Policy, and Management at the Boston University School of Public Health in August 2017. Her work focuses on applying econometric methods to evaluate health and social policies and has received national and international news coverage, including from *The New York Times*, *The Guardian*, and *National Public Radio*. Dr Raifman earned her doctoral and Masters degrees from the Department of Global Health and Population at the Harvard T.H. Chan School of Public Health.

Felix Lam is the Senior Manager of Monitoring and Evaluation at the Clinton Health Access Initiative. Mr Lam is leading evaluations on scaling up the use of ORS and zinc in India, Nigeria, Uganda, and Kenya. Mr Lam has a Master of Science from the Harvard T.H. Chan School of Public Health and a Bachelor of Arts from the University of California, Berkeley.

Janeen Madan Keller is a policy analyst at the Center for Global Development, focusing on global health issues. Prior to joining CGD in June 2015, she spent 2 years in Dakar, Senegal, where she worked with the UN World Food Program on nutrition and food security programs across 18 countries in West Africa. She has also worked with UNICEF in Bamako, Mali and conducted research on changing health behaviours in Niger. Originally from Mumbai, India, she holds an MS in public health nutrition and food security from Tufts University and a BA with honours in political science and French from Vassar College.

Alexander Radunsky is a doctoral candidate at the Harvard T.H. Chan School of Public Health. He specialises in qualitative and mixed methods analyses and is completing his dissertation research in Kwa-Zulu Natal, South Africa.

William Savedoff is a senior fellow at the Center for Global Development where he works on issues of aid effectiveness and health policy. His current research focuses on the use of performance payments in aid programmes and problems posed by corruption. At the Center, Savedoff played a leading role in the Evaluation Gap Initiative and coauthored Cash on Delivery Aid with Nancy Birdsall. Before joining the Center, Savedoff prepared, coordinated, and advised development projects in Latin America, Africa and Asia for the Inter-American Development Bank and the World Health Organization. As a Senior Partner at Social Insight, Savedoff worked for clients including the National Institutes of Health, Transparency International, and the World Bank. He has published books and articles on labour markets, health, education, water, and housing including 'What Should a Country Spend on Health?', 'Governing Mandatory Health Insurance, and Diagnosis Corruption.

evaluations and less than 10 per cent of performance evaluations used relevant data, followed accepted social science methods for sampling, or had high analytical validity and reliability. There is a need to improve the methodological quality of programme evaluations.

Keywords

Evaluation; methodology; global health; development

Introduction

In 2015, official development assistance (ODA) was nearly US\$191 billion (Organization for Economic Cooperation and Development 2017). While ODA is associated with significant improvements in low- and middle-income countries, it is also under scrutiny for its effectiveness (Craig and Dollar 1997; Easterly, Levine, and Roodman 2003; Glassman and Temin 2016). Evaluations play a critical role in ODA programmes by providing accountability for the use of funds and by contributing learning on how to design and implement more effective programmes.

ODA has always been subject to some evaluation, but more recently, a strong international consensus has emerged over the need for more and better evaluation. In 2005, 91 countries endorsed the Paris Declaration on Aid Effectiveness (“Paris Declaration on Aid Effectiveness” 2005), which aimed to improve the impact of ODA through, among other things, better measurement of programme results and improved accountability for achieving goals. The following year, the Center for Global Development’s Evaluation Gap Working Group (EGWG) released a report highlighting a lack of knowledge regarding aid programmes and calling on governments and ODA funders to increase funding for impact evaluations, to strengthen monitoring and evaluation systems, to improve standards for evidence, and to facilitate access to knowledge (Evaluation Gap Working Group 2006).

Since the Paris Declaration and the EGWG report, the evaluation landscape has evolved. The International Initiative for Impact Evaluations (3ie) was created in 2008, exclusively to fund evaluations of development policies and programmes in low- and middle-income countries (3ie: International Initiative for Impact Evaluation 2015). The largest ODA funders have developed new evaluation policies (“DFID Evaluation Policy 2013 – Publications – GOV.UK”; U.S. Agency for International Development 2016b; Bill and Melinda Gates Foundation 2017) and are implementing more programme evaluations (Kennedy-Chouane and Lundgren n. d.; Independent Commission for Aid Impact 2014; U.S. Agency for International Development 2016a).

While the number of evaluations has increased, we know less about their quality and use. An independent meta-analysis of the United States Agency for International Development’s (USAID) evaluations showed substantial room for improvement. It found that the methods used in these evaluations were based on specific evaluation questions in just 22 per cent of the cases published in 2012; and only 34 per cent of evaluations contained descriptions of analytical methods (“Meta-Evaluation of Quality and Coverage of USAID Evaluations”). These findings are mirrored in a Government Accountability Office (GAO) study of foreign

aid evaluations conducted in 2015 by six US government agencies – USAID, the Millennium Challenge Cooperation, Department of Agriculture, Department of Defense, Department of State, and Department of Health and Human Services. The GAO study assessed evaluations against three broad quality criteria including design, implementation, and analysis. The review found that only about a quarter of evaluations met all quality criteria and an additional 50 per cent partially met these criteria. The study underscored the need to improve evaluation quality, especially in the areas of sampling, data collection, and analysis (Government Accountability Office 2017).

Furthermore, evaluations may not be widely used. A review of World Bank policy reports, which included evaluations, indicated that 31 per cent of reports had never been downloaded (Doemeland and Trevino 2014). The 2016 OECD DAC Peer Review of the United States also suggests that better planning and management would make evaluations more effective in promoting learning across USAID (Organization for Economic Cooperation and Development 2016a).

In this study, we assessed the extent to which ODA evaluations commissioned or conducted by major health funders adhered to social science methodological standards. We followed the King, Keohane, and Verba (1994) philosophy that evaluations have the same underlying logic regardless of their size or scope or whether they are quantitative or qualitative. Both quantitative and qualitative approaches involve using systematic data collection with the aim of making descriptive or explanatory inferences about the world. The extent to which evaluators take a systematic approach affects the relevance, validity, and reliability of either quantitative or qualitative data. We intentionally choose these three criteria for assessing evaluation quality because they readily encompass the range of purposes and characteristics sought by evaluators with different perspectives on evaluation (Rossi, Lipsey, and Freeman 2003; Campbell 1969; Cronbach and Shapiro 1982; Patton 1997). We applied these criteria to both quantitative and qualitative evaluations.

Methods

Sample

To assess the quality of evaluations, we selected a sample of global health programme evaluations from those commissioned or conducted by major DAH funders, including both impact and performance evaluations. We chose to focus on evaluations of global health programmes because health comprises one of the largest components of ODA (14 per cent in 2015)¹ and because the global health field is widely perceived as a pioneer in conducting evaluations (Cameron, Mishra, and Brown 2016). We also focused on publicly available evaluations of large-scale programmes implemented to improve public health, excluding small-scale evaluations of programmes implemented for the purpose of conducting a study. We included evaluations published between 2009 and 2014, capturing a period of 5 years prior to the year when the study began in 2015.

¹Data accessed on 7 June 2017 from stats.oecd.org; figures represent commitments in current US\$ and represent the following two categories: Health (1.2) and Population Policies/Programmes and Reproductive Health (1.3).

To further refine the universe of evaluations, we focused on five major funders of DAH based on data from the OECD: USAID, the Global Fund to Fight AIDS, Tuberculosis and Malaria (Global Fund), the President's Emergency Fund for AIDS Relief (PEPFAR), the United Kingdom's Department for International Development (DFID), and the International Development Association (IDA) at the World Bank. These five funders committed about US \$18.2 billion in health aid to low- and middle-income countries in 2015 (OECD, 2017). We initially sought to include the Bill and Melinda Gates Foundation (BMGF) in the study, but dropped them from our sample after BMGF staff indicated that most BMGF funding typically goes towards small-scale studies of health interventions, rather than towards implementing large-scale health programmes.

The evaluations from which we sampled included evaluations that these funders funded to evaluate their programmes, and covered a broad variety of programmes and projects that provided direct services, in-kind medications, budget support, medical training, logistics support, infrastructure, and loans for policy reforms. For the five funding agencies included in our sample, we generated a list of publicly available evaluations on health and water and sanitation programmes through searches on organisation websites, Google, Google Scholar, and the 3ie Impact Evaluation Database. We also contacted the heads of evaluation departments at each organisation to confirm that our list captured all relevant evaluations. All of the evaluations were published on funders' websites, and we did not identify any articles in peer-reviewed journals that met our inclusion criteria. The universe of evaluations was finalised on 1 September 2015.

We included all evaluations regardless of their methodological approach. We stratified our sample, classifying evaluations as either impact or performance evaluations. We defined 'impact evaluations' as 'attribution analyses' (White 2010), or evaluations whose primary purpose was to measure the impact that could be attributed to a particular intervention. Impact evaluations typically involve quantitative analyses involving randomised trials, quasi-experimental designs, or regression analyses. We used the term 'performance evaluations' to refer to evaluations that focused on other aspects of performance such as managerial efficiency, outputs, or beneficiary satisfaction. Performance evaluations typically use qualitative methods. From the list of 299 evaluations, we randomly chose five impact evaluations and five performance evaluations from each funder. In some cases, funders had fewer than five of either type of evaluation (for example, the Global Fund and PEPFAR), so we assessed all available evaluations in that category. Our final sample comprised 37 evaluations from 5 funders.

Assessment instrument

Seven evaluators trained in evaluative methods reviewed the sample. We developed the assessment instrument based on social science standards for research (Maxwell 2004; Patton 1990; King, Keohane, and Verba 1994) and on the Cochrane rubric for assessing risk of bias (Higgins and Green 2011). Similarly to the Cochrane rubric, questions about whether each methodological area was described in enough detail to assess quality were followed by questions about methodological quality. We tested the assessment instrument by having all

reviewers assess an initial set of 11 evaluations and by revising the instrument based on that review to ensure that the questions and concepts were clear.

Once we were satisfied with the instrument, two evaluators independently completed the assessment form provided in Appendix 1 for each evaluation in the sample. Then, evaluators met to discuss and resolve any discrepancies in their assessments. During this meeting, evaluators filled out the assessment form again to indicate their final answers on measures on which they initially had discordant answers.

The main outcomes of interest were: (1) The proportion of all evaluations that were impact evaluations relative to performance evaluations; and (2) the relevance, validity, and reliability of evaluations. We defined evaluations as *relevant* if the authors addressed questions related to the means or ends of a programme or intervention, and used pertinent data to answer those questions. We defined evaluations as *valid* if analyses were methodologically sound and conclusions derived logically and consistently from the findings. We defined evaluations as *reliable* to the extent that the conclusions would likely be confirmed if the evaluation was repeated in a similar context. We assessed validity and reliability for both sampling and analysis, respectively. Other outcomes of interest included ethical conduct and conflict of interest considerations. Across each primary and secondary outcome, we ranked evaluations as having low, medium, or high scores by aggregating data from the assessment form, as indicated in Appendix 2.

Analysis

We reported descriptive statistics on the number of publicly available evaluations from each organisation relative to their financial contributions to DAH. We also reported the aggregate main outcome measures, distinguishing between impact and performance evaluations. As detailed in Appendix 2, we developed the main outcome measures based on a three-point scale, with three being the best score. We compiled scores based on the rationale that a low or medium score on one component could undermine that entire aspect of the analysis. An evaluation received a three if all component criteria received scores of three, a two if one of the criteria scored a two, and a one if one or more components scored a one or multiple components scored twos.

To assess the reliability of our assessments, we calculated the per cent agreement for each outcome and its components (Di Eugenio and Glass 2004; Viera and Garrett 2005). We decided to use per cent agreement rather than Cohen's kappa, which is based on the assumption that reviewers' responses are random (McHugh 2012). We used linear regression analyses to assess whether prior planning of evaluations was associated with two outcomes – sampling validity and reliability and analytical validity and reliability – controlling for evaluation type and using robust standard errors. We planned to assess the relationship between evaluation and programme costs and quality but were unable to obtain precise data on the costs of evaluations or the programmes they evaluated. We did not aim to assess whether impact or performance evaluations were most appropriate for each evaluation.

Findings

The US Government provides the greatest amount of DAH through USAID and PEPFAR, averaging US\$7.4 billion per year between 2010 and 2012 (see Table 1). USAID had the most publicly available evaluations within the 2009 to 2014 period, with 254 evaluations, and PEPFAR had 16 evaluations. The Global Fund provided US\$2.7 billion in DAH on average and had four evaluations. DFID provided US\$1.5 billion and had 14 evaluations. The World Bank provided US\$1.1 billion and had 24 evaluations. Most evaluations (91%) were performance evaluations, 7 per cent were impact evaluations, and 2 per cent were a combination. Randomised controlled trials (RCTs) made up less than 2 per cent of all evaluations in the universe and 23 per cent of all impact evaluations (Figure 1). USAID had the smallest share of impact evaluations (4%), while the World Bank had the greatest share (30%).

In the second panel of Table 1, we describe the 37 evaluations in our sample by funder and type. We also list each evaluation in our sample in Supplementary Appendix 3. Our sample had 21 performance and 13 impact evaluations, and three that were a combination of both. Our final sample differed slightly from our initial sampling strategy because during analysis, we found a few evaluations were initially misclassified. For the three evaluations in our sample that were a combination of impact and performance evaluations, we based our analysis on the ‘impact’ sections, and subsequently treated these as impact evaluations for the purposes of this study.

We present the distribution of relevance, reliability, and validity scores in Table 2. We first assessed whether the objectives of the evaluation reflected the main objectives of the health programme being evaluated. Most evaluations had relevant objectives, with 88 per cent of impact evaluations and 67 per cent of performance evaluations having highly relevant objectives. Second, we assessed the relevance of data by considering whether the primary or secondary data used for evaluations reflected the population being served or services being provided. We found that 38 per cent of impact evaluations and no performance evaluations used highly relevant data, while 67 percent of performance evaluations and 13 per cent of impact evaluations used data of low relevance. We found only four performance evaluations collected data from programme recipients and nonrecipients. Moreover, only two performance evaluations used a methodologically rigorous analysis approach. Five of the 13 performance evaluations that interviewed programme recipients selected respondents through convenience sampling and were unable to address potential for bias. Overall, we found that 38 per cent of impact evaluations and 10 per cent of performance evaluations had high sampling validity. Most performance evaluations (81%) and half of the impact evaluations (50%) had low sampling validity. Finally, we assessed analytical validity based on whether evaluations took a systematic approach to the analysis and addressed or considered potential confounders. We found that about a third (31%) of impact evaluations and 5 per cent of performance evaluations had high analytical validity. Most performance evaluations (71%) had low analytical validity, while half of impact evaluations had medium validity (50%).

We present the per cent agreement for each measure in Table 3. Agreement between reviewers exceeded 80 per cent for each outcome, ranging from 83.1 per cent for the relevance of data to 94.3 per cent on the relevance of objectives.

We also investigated the association between prior planning of evaluations, the use of baseline data, and the use of monitoring data with the scores for relevant objectives, sampling validity, data relevance, and addressing confounders using linear regression analyses. Evaluations were planned prior to programme implementation for 9 (24%) evaluations; only one of these was a performance evaluation. As conveyed in Table 4, prior planning of impact evaluations was positively associated with increased use of relevant data (2.23, 95% CI: 1.30–3.82) and better analytical validity and reliability (2.67, 95% CI: 1.57–4.53), relative to impact evaluations that were not planned prior to programme implementation. Use of baseline data was mentioned in 5 per cent of performance evaluations and 63 per cent of impact evaluations. Use of monitoring data was mentioned by 76 per cent of performance evaluations and 50 per cent of impact evaluations. In our analysis, use of baseline or monitoring data was not associated with evaluation quality.

Secondary outcomes of interest included consideration of ethics and data transparency. In 88 per cent of evaluations, we could not identify any clear conflicts of interest for the evaluators, though no evaluations included statements that explicitly indicated there *were* no conflicts of interest. We found that 17 (46%) of the evaluations mentioned ethics considerations in data collection and that only 8 (22%) explicitly mentioned receiving institutional review board approval.

The data collected in this study are publicly available at www.cgdev.org.

Discussion

Global health programme evaluations are key to learning about the best approaches for scaling up and delivering health programmes. Evaluations are also critical for accountability to the taxpayers who provide billions of dollars in development funding and to the citizens who should be served by global health programmes. We found that most evaluations were performance evaluations rather than impact evaluations; among impact evaluations, RCTs constitute a minority. In addition, fewer than half of impact evaluations and fewer than 10 per cent of performance evaluations met social science standards for relevance, validity, and reliability. We found that the relevance of objectives and data were greater for impact evaluations than performance evaluations. Overall, our findings suggest the need to improve the relevance, validity, and reliability of global health programme evaluations to better facilitate learning and accountability.

Prior planning

One of the most basic ways to improve evaluation quality involves early and better planning. Only nine evaluations (of which only one was a performance evaluation) referenced evaluation planning prior to programme implementation. The remaining evaluations lacked references to pre-programme planning. Nevertheless, prior planning was associated with the use of more relevant data and better analytical validity. Performance evaluations often

included interviews with project implementers, but only 62 per cent (13) of them interviewed programme beneficiaries. Of those evaluations that interviewed beneficiaries, 40 per cent used convenience sampling, which can easily lead to biased findings. Furthermore, for 18 performance evaluations in our sample, reviewers thought programme nonrecipients should have been sampled as well to enhance the validity of the findings. However, just four evaluations (22%) included interviews with programme nonrecipients. Performance evaluations could be improved through more deliberate and preferably representative sampling of both programme recipients and nonrecipients.

Moving away from an evaluation dichotomy

Recent debates on methodological approaches to evaluating development economics questions have focused on the merits and drawbacks of RCTs (Cohen and Easterly 2010), a debate that has carried over to the programme evaluation community (Brass, Nunez-Neto, and Williams 2008). In our review of programme evaluations by five major funders of DAH, we found that the spectrum of evaluation types is heavily weighted towards performance evaluations. Less than 2 per cent (5) of evaluations were RCTs. In contrast, 92 per cent of evaluations were performance evaluations, most of which used qualitative methods. These performance evaluations had lower scores on the relevance of data, sampling validity and reliability, and analytical validity and reliability.

Impact evaluations are achieving better methodological rigour even though deficiencies persist. By contrast, methodological rigour in performance evaluations is lacking. Given that there are so many performance evaluations conducted, it is essential to improve the quality of these evaluations by encouraging less biased approaches to collecting information and greater rigour and replicability in the methods used to derive conclusions. Funders should also carefully consider which methods best answer evaluation questions of interest.

Evaluator independence and ethics

Our sample is made up of evaluations commissioned or conducted by funders to assess their own projects, which can encourage positive bias (White and Bamberger 2008). Some of these evaluations were contracted by operational staff, but most came from offices that are formally separated from operational departments to mitigate this problem. For example, the World Bank's Independent Evaluation Group reports directly to the institution's Board of Directors, not the President; Britain created the Independent Commission for Aid Impact, which reports directly to Parliament and not to DFID. In many cases, external evaluators were contracted from research institutions or consulting firms.

As White and Bamberger (2008) note, we have no evidence that particular arrangements are more or less subject to positive bias, and it is encouraging that reviewers judged evaluators to lack independence in only 16 per cent of the cases (6 of the 37 evaluations). However, the credibility of evaluations would be enhanced with greater attention to positive bias, particularly by disclosing institutional affiliations, referring to applicable professional standards, statements on conflicts of interest, and identifying contributions by implementers in sampling, data collection, or writing.

All the evaluations gathered information from a range of individuals, which raises potential ethical issues. For example, interviews with individuals who have socially stigmatised illnesses must be conducted appropriately to protect their rights. Nevertheless, less than half of the evaluations mentioned ethics considerations in data collection, and less than one quarter mentioned institutional review board approval. Ensuring confidentiality and preventing undue harm to evaluation participants has not become a standard practice among DAH evaluations.

Strengths and limitations

This study has several strengths and limitations. Our approach to assessing evaluation quality appears robust. The process of jointly developing the protocol and testing led to clearer definitions and ultimately to substantial agreement between reviewers even though scores were assigned independently. Our findings are more generalisable than other assessments of evaluation because we selected a representative sample from a well-defined universe of evaluations. To our knowledge, it is also the first assessment of the quality of programme evaluations to encompass different funders. Our method of assigning multiple reviewers and testing instruments generated consistent findings which increases the credibility of the results. However, the large amount of time it took to assess each evaluation led to a relatively small sample, only 37 evaluations. This limited our ability to identify patterns, compare funders, and differentiate among methods. We were also unable to assess the extent to which evaluations were used to inform implementation of subsequent programmes.

Recommendations for future evaluation assessments

Conducting the study gave us insight into considerations for similar assessments of evaluations in the future, particularly if funders or other organisations are able to access additional data. In developing the sample, we only had access to publicly available evaluations and did not have access to all of the evaluations that funders commissioned or conducted. We also would have ideally liked to prioritise the health programmes that represented the greatest monetary investment to assess the methodological quality of those evaluations. Funders may have data on expenditures that could inform a sampling strategy based on cost and other metrics. Similarly, funders may have more data on whether or not evaluations were planned from the outset, on the existence of monitoring and evaluation data, and on evaluators' access to and use of monitoring and evaluation data. Finally, funders may have access to data on downloads of evaluation reports to assess readership of evaluations such as that conducted by Doemeland and Trevino (2014) at the World Bank.

In addition to next steps that are possible for funders, we could have better distinguished between different types of monitoring and evaluation data. While we observed that many evaluations did not report any quantitative data, we did not distinguish between whether evaluations made use of quantitative monitoring and evaluation data or used monitoring approaches such as document review. To inform the role of monitoring and evaluation data in strengthening evaluation methodology, future assessments could collect more specific information about what data were used and how.

Conclusion: achieving learning and accountability with global health programme evaluations

Evaluations are critical to learning about what works in global development and for holding funders and implementers accountable. We reviewed evaluations from the five large DAH funders, which together commit between US\$12–18 billion to health programmes annually. A small share of these evaluations met social science standards, demonstrating that good quality evaluations can be conducted and contribute to learning and accountability. However, agencies must increase this share of high-quality evaluations to assure the relevance, validity, and reliability of findings. A key recommendation from our analysis is to plan more evaluations before programme implementation begins. We also call for the global development community to have an intensive discussion about the methodological quality of performance evaluations, which make up 92 per cent of publicly available evaluations from five large DAH funders, and about the need to better invest in collecting and using robust and varied types of data. Finally, funders and programme evaluators should more explicitly address conflicts of interest and follow appropriate standards for protecting human subjects. Taking these steps to improve the quality of evaluations is critical to ensuring that global health interventions – and development programmes more generally – effectively and efficiently contribute to improved lives and increased wellbeing around the world.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors would like to thank Victoria Fan, Rifaiyat Mahbub, Carmel Salhi, and Jesse Heitner for their contributions to the study. This work was supported by the National Institute of Allergy and Infectious Diseases grant T32AI102623.

Disclosure statement

No potential conflict of interest was reported by the authors.

References

- 3ie:International Initiative for Impact Evaluation. 3ie Annual Report. 2015. <http://www.3ieimpact.org/>
- Bill and Melinda Gates Foundation. Evaluation Policy 2017. 2017. <http://www.gatesfoundation.org/How-We-Work/General-Information/Evaluation-Policy>
- Brass CT, Nunez-Neto B, Williams ED. Congress and Program Evaluation: An Overview of Randomized Controlled Trials (Rcts) and Related Issues. Hauppauge, NY: Nova Science Publishers; 2008. http://digital.library.unt.edu/ark:/67531/metacrs9145/m1/1/high_res_d/RL33301_2006Mar07.pdf
- Cameron DB, Mishra A, Brown AN. The Growth of Impact Evaluation for International Development: How Much Have We Learned? Journal of Development Effectiveness. 2016; 8(1):1–21. DOI: 10.1080/19439342.2015.1034156
- Campbell DT. Reforms as Experiments. American Psychologist. 1969; 24(4):409.doi: 10.1037/h0027982
- Cohen J, Easterly W. What Works in Development?: Thinking Big and Thinking Small. Brookings Institution Press; 2010. <https://books.google.com/books?>

hl=en&id=RMvD8HuWfegC&oi=fnd&pg=PP1&dq=what+works+in+development+thinking+big+thinking+small&ots=97N0zqvPli&sig=pt2yZzFJkjtjTbedR_5gux3FQ3I

Craig BA, Dollar D. Aid, Policies, and Growth. 1997. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=569252

Cronbach LJ, Shapiro K. , et al. Designing Evaluations of Educational and Social Programs. Jossey-Bass; 1982. <http://agris.fao.org/agris-search/search.do?recordID=US201300437175>

[Accessed January 19, 2017] DFID Evaluation Policy 2013 – Publications – GOV.UK. n.d. <https://www.gov.uk/government/publications/dfid-evaluation-policy-2013>

Di Eugenio B, Glass M. The Kappa Statistic: A Second Look. Computational Linguistics. 2004; 30(1): 95–101. DOI: 10.1162/089120104773633402

Doemeland D, Trevino J. Policy Research Working Paper 6851: Which World Bank Reports Are Widely Read?. 2014. <http://documents.worldbank.org/curated/en/387501468322733597/pdf/WPS6851.pdf>

Easterly W, Levine R, Roodman D. New Data, New Doubts: Revisiting 'aid, Policies, and Growth'. 2003. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=425001

Evaluation Gap Working Group. When Will We Ever Learn? Improving Lives Through Impact Evaluation. Center For Global Development. 2006. <http://cgdev.org.488elwb02.blackmesh.com/publication/when-will-we-ever-learn-improving-lives-through-impact-evaluation>

Glassman A, Temin M. Millions Saved: New Cases of Proven Success in Global Health. Washington, DC: Brookings Institution Press; 2016.

Government Accountability Office. Report to Congressional Requesters, GAO-17-316: Foreign Assistance: Agencies Can Improve the Quality and Dissemination of Program Evaluations. 2017. <http://www.gao.gov/assets/690/683157.pdf>

Higgins JPT, Green S. Cochrane Handbook for Systematic Reviews of Interventions. Vol. 4. John Wiley & Sons; 2011. <https://books.google.com/books?hl=en&id=NKMg9sMM6GUC&oi=fnd&pg=PT13&dq=Cochrane+Handbook+for+Systematic+Reviews+of+Interventions+&ots=LHZGLYGGb1&sig=IQwXahzWltQMjIORi8jW8jGf93Q>

Independent Commission for Aid Impact. How DFID Learns. Report 34. 2014. <http://icai.independent.gov.uk/wpcontent/uploads/How-DFID-Learns-FINAL.pdf>

Kennedy-Chouane M, Lundgren H. Evaluating Development Activities – 12 Lessons from the OECD DAC. Paris, France: OECD; Evaluating Development Activities. n.d. http://www.oecd.org/dac/peer-reviews/12_Less_eval_web.pdf

King G, Keohane RO, Verba S. Designing Social Inquiry: Scientific Inference in Qualitative Research. Princeton university press; 1994. <https://books.google.com/books?hl=en&id=A7VFF-JR3b8C&oi=fnd&pg=PP1&dq=methods+king+keohane&ots=f1muT7zPZV&sig=K0IpegP-4Pau7dJnPy0g7184jHw>

Maxwell JA. Using Qualitative Methods for Causal Explanation. Field Methods. 2004; 16(3):243–264. DOI: 10.1177/1525822X04266831

McHugh ML. Interrater Reliability: The Kappa Statistic. Biochemia Medica. 2012; 22(3):276–282. DOI: 10.11613/BM.2012.031 [PubMed: 23092060]

[Accessed January 19, 2017] Meta-Evaluation of Quality and Coverage of USAID Evaluations. <https://www.usaid.gov/evaluation/meta-evaluation-quality-and-coverage>

Organization for Economic Cooperation and Development. [Accessed June 7, 2017] Creditor Reporting System. 2017. <https://stats.oecd.org/Index.aspx?DataSetCode=CRS1>

Paris Declaration on Aid Effectiveness. 2005. <https://www.oecd.org/dac/effectiveness/34428351.pdf>

Patton MQ. Qualitative Evaluation and Research Methods. Thousand Oaks, CA: SAGE Publications; 1990. <http://psycnet.apa.org/psycinfo/1990-97369-000>

Patton MQ. Utilisation-Focused Evaluation: The New Century Text. Utilization Focused Evaluation: The News Century Text. 1997. <http://www.alnap.org/resource/11131>

Rossi PH, Lipsey MW, Freeman HE. Evaluation: A Systematic Approach. Thousand Oaks, CA: SAGE Publications; 2003. <https://books.google.com/books?hl=en&id=QF9WBAAAQBAJ&oi=fnd&pg=PT10&dq=Evaluation:+A+Systematic+Approach+rossi&ots=9zw86FaOF4&sig=g4T6pN-AFXer-rBSTKM05sY8jN4>

- U.S. Agency for International Development. Five Years of Evaluation Practice. Washington, D.C: 2016a. [https://www.usaid.gov/sites/default/files/documents/1870/Strengthening Evidence-Based Development – Five Years of Better Evaluation Practice at USAID.pdf](https://www.usaid.gov/sites/default/files/documents/1870/Strengthening%20Evidence-Based%20Development%20%E2%80%93%20Five%20Years%20of%20Better%20Evaluation%20Practice%20at%20USAID.pdf)
- U.S. Agency for International Development. USAID Evaluation Policy. Washington, D.C: 2016b. <https://www.usaid.gov/sites/default/files/documents/1870/USAIDEvaluationPolicy.pdf>
- Viera AJ, Garrett JM, et al. Understanding Interobserver Agreement: The Kappa Statistic. Family medicine. 2005; 37(5):360–363. [PubMed: 15883903]
- White H. A Contribution to Current Debates in Impact Evaluation. Vol. 16. Thousand Oaks, CA: SAGE Publications; 2010.

Figure 1. All evaluations in universe by type and methodology

Source: Authors' calculations.

Notes: DiD: difference in differences; RCT: randomised controlled trial. The categorisation presented in this figure differs slightly from Table 1. The total number of performance evaluations is 275, this includes three evaluations which were initially coded as 'impact' in our universe and turned out to be 'performance'. Note also that three evaluations that were categorised as 'performance & impact' were included in our sample and treated as 'impact' evaluations – two used DiD methods and 1 used quasi-experimental methods.

Table 1

Type of evaluation: universe versus sample: 2009–2014.

Funder	Average annual DAH 2010–12 (billions)	Universe				Sample			
		Total	Impact	Performance	Both	Total	Impact	Performance	Both
		N	N (%)	N (%)	N (%)	N	N (%)	N (%)	N (%)
USAID	7.4	239	10 (4)	226 (95)	3 (1)	10	3 (30)	6 (60)	1 (10)
PEPFAR		19	2 (11)	16 (84)	1 (5)	5	1 (20)	4 (80)	–
Global Fund	2.7	4	1 (25)	3 (75)	–	3	2 (67)	1 (33)	–
DFID	1.5	19	2 (14)	11 (79)	1 (7)	10	3 (30)	5 (50)	2 (20)
World Bank	1.1	23	7 (30)	16 (70)	–	9	4 (44)	5 (56)	–
Total	12.7	299	22 (7)	272 (91)	5 (2)	37	13 (35)	21 (57)	3 (8)

Source: Figures from average annual DAH are from OECD’s Creditor Reporting System. Accessed 30 January 2017.

Note: In subsequent analyses, we treated the three evaluations in our sample labelled as ‘both’ as ‘impact’ evaluations; reviewers were asked to focus on the ‘impact’ evaluation sections of those evaluations.

Table 2

Summary scores for evaluations in sample ($N = 37$).

Scores	Evaluation type	Low <i>N</i> (%)	Medium <i>N</i> (%)	High <i>N</i> (%)
Relevance of objectives	Impact ($N = 16$)	0 0%	2 13%	14 88%
	Performance ($N = 21$)	1 5%	6 29%	14 67%
Relevance of data	Impact ($N = 16$)	2 13%	8 50%	6 38%
	Performance ($N = 21$)	14 67%	7 33%	0 0%
Sampling validity & reliability	Impact ($N = 16$)	8 50%	2 13%	6 38%
	Performance ($N = 21$)	17 81%	2 10%	2 10%
Analytical validity & reliability	Impact ($N = 16$)	3 19%	8 50%	5 31%
	Performance ($N = 21$)	15 71%	5 24%	1 5%

Percentages may add to more than 100 per cent due to rounding

Table 3

Inter-rater reliability on the relevance, validity, and reliability scores ($N = 37$).

Score	Percent agreement
Relevance of objectives	94.3
Relevance of data	83.1
Sampling validity and reliability	85.3
Analytical validity and reliability	88.9

Table 4

Associations between evaluation characteristics and quality measures of relevance, reliability, and validity.

Characteristic	Relevance of objectives	Relevance of data	Sampling validity and reliability	Analytical validity and reliability
Prior planning	0.67 (0.42–1.07)	2.23 *** (1.30–3.82)	1.60 (0.54–4.73)	2.67 *** (1.57–4.53)
Use of baseline data	0.60 (0.35–1.05)	1.03 (0.57–1.84)	1.23 (0.57–2.66)	1.09 (0.60–1.99)
Use of monitoring data	0.89 (0.66–1.21)	1.12 (0.73–1.72)	0.63 (0.32–1.26)	0.74 (0.43–1.27)

We conducted linear regression analyses controlling for evaluation type in these analyses and results use robust standard errors.

<0.01.