Evaluating Evaluations:

Assessing the Quality of Aid Agency Evaluations in Global Health

Julia Goldberg Raifman, Felix Lam, Janeen Madan Keller, Alexander Radunsky, and William Savedoff

Abstract

Evaluations are key to learning and accountability yet the quality of those evaluations are critical to their usefulness. We assessed the methodological quality of global health program evaluations commissioned or conducted by five major funders and published between 2009 and 2014. From a universe of 299 large-scale global health program evaluations, we randomly selected 37 evaluations stratified by whether they were performance evaluations or impact evaluations and applied a systematic assessment approach with two reviewers scoring each evaluation. We found that most evaluations did not meet social science methodological standards in terms of using methods and data that would simultaneously assure relevance, validity, and reliability.

Most evaluations (76 percent) asked questions relevant to the health program, but 43 percent of evaluations failed to collect relevant data. In addition, only about a fifth of the evaluations followed accepted social science methods for sampling. We also assessed whether evaluations took a systematic analytical approach and considered potential confounding variables. In this regard, only 16 percent of evaluations in our sample had high analytical validity and reliability.

The study provides ten recommendations for improving the quality of evaluations, including a robust finding that early planning of evaluations is associated with better quality and noting the value of better sampling approaches in data collection and disclosure of potential conflicts of interest and data.

Keywords: evaluation, impact evaluation, health, foreign aid, learning, research quality

JEL Codes: D04, F35, F53, H43, I18

Center for Global Development
www.cgdev.org

Working Paper 461
August 2017

Evaluating Evaluations: Assessing the Quality of Aid Agency Evaluations in Global Health

Julia Goldberg Raifman
Boston University

Felix Lam
Clinton Health Access Initiative

Janeen Madan Keller
Center for Global Development

Alexander Radunsky
Harvard T.H. Chan School of Public Health

William Savedoff
Center for Global Development

We would like to thank Victoria Fan, Rifaiyat Mahbub, Carmel Salhi, and Jesse Heitner for their contributions to the study. We also appreciate the comments we received from Sebastian Bauhoff, agency staff from the United States Agency for International Development, the Global Fund to Fight AIDS, Tuberculosis, and Malaria, the World Bank, and the UK Department for International Development, and one anonymous reviewer.

The Center for Global Development is grateful for contributions from the Lakeshore Foundation and the William and Flora Hewlett Foundation in support of this work.

Julia Goldberg Raifman, 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. https://www.cgdev.org/publication/evaluating-evaluations-assessing-quality-aid-agency-evaluations-global-health

Center for Global Development
2055 L Street NW
Washington, DC 20036

202.416.4000
(f) 202.416.4050

www.cgdev.org

Contents

Conflict of interest

None of the authors received funding from any of the agencies that were reviewed to directly support this study. The Center for Global Development, which provided staff time for this study, has received grants from DFID and held contracts with the World Bank for other activities.

Time and cost

This study began in 2014 and was completed in 2017. We estimate that the team spent approximately 12 full-time equivalent weeks to review the 37 evaluations and that another 18 full-time equivalent weeks were spent on analysis, team discussion, write-up, and revisions. We cannot provide a cost estimate since the project was not supported by a direct grant.

Ethics and transparency

This study did not involve human subjects. The agencies whose evaluations were being assessed were contacted in January 2014 and asked for comments on the research protocol and to add or amend the list of evaluations in the universe. They were also given opportunities to comment on a previous draft of this paper. The data and computer code for analysis are publicly available at https://www.cgdev.org/media/assessing-quality-global-health-evaluations.

Acronyms

3ie	International Initiative for Impact Evaluations
BMGF	Bill and Melinda Gates Foundation
DAC	Development Assistance Committee
DAH	Development assistance for health
DFID	Department for International Development
EGWG	CGD's Evaluation Gap Working Group
GAO	Government Accountability Office
Global Fund	The Global Fund to Fight AIDS, Tuberculosis, and Malaria
ICAI	Independent Commission for Aid Impact
IEG	World Bank's Independent Evaluation Group
ODA	Oversees Development Assistance
OECD	Organization for Economic Cooperation and Development
PEPFAR	President's Emergency Fund for AIDS Relief
RCT	Randomized Controlled Trial
USAID	United States Agency for International Development

Introduction

In 2015, Official Development Assistance (ODA) amounted to US$191 billion (OECD DAC 2017). While ODA is associated with significant improvements in social sector outcomes in low- and middle-income countries, it is also under scrutiny for its effectiveness (Burnside and Dollar 2000; Easterly et al. 2004; Glassman and Temin 2016). Evaluations play a critical role in ODA programs by providing accountability for the use of funds and by informing how to design and implement more effective programs.

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 (OECD DAC 2005), which aimed to improve the impact of ODA through, among other things, better measurement of program 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 programs and calling on governments and funders to increase funding for impact evaluations, to strengthen monitoring and evaluation systems, to improve standards for evidence, and to facilitate access to knowledge (EGWG 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 programs in low- and middle-income countries (3ie 2015). Furthermore, the largest ODA funders have developed new evaluation policies—the United Kingdom Department for International Development (DFID) published its evaluation policy in 2013; the United States Agency for International Development (USAID) in 2011 and updated in 2016; and the Bill and Melinda Gates Foundation (BMGF) in 2017—and are conducting and commissioning more program evaluations (IEG 2012; Kennedy-Chouane and Lundgren 2013; ICAI 2014; USAID 2016).

While the number of evaluations has increased, we know less about their quality and use. An independent meta-analysis of USAID's evaluations published in 2012 showed substantial room for improvement. It found that the methods used in evaluations were based on specific evaluation questions in just 22 percent of the cases; and only 34 percent of evaluations contained descriptions of analytical methods (Hageboeck et al. 2013). These findings are mirrored in a recent Government Accountability Office (GAO) study of foreign aid evaluations completed in fiscal year 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 percent partially met these criteria. The study underscored the need to improve evaluation quality, especially in the areas of sampling, data collection, and analysis (GAO 2017).

Furthermore, evaluations are not widely used, and management and planning can be weak. Within the World Bank for example, 31 percent of policy reports, which include evaluations, had never been downloaded (Doemeland and Trevino 2014). In addition, the 2016 OECD Development Assistance Committee (DAC) Peer Review of the United States commends USAID for building a culture of evaluation, but argues that better planning and management would make evaluations more effective in promoting learning across the agency (OECD DAC 2016, chapter 6).

This study aimed to assess the quality of ODA evaluations commissioned or conducted by major health funders—USAID, the Global Fund to Fight AIDS, Tuberculosis, and Malaria (the Global Fund), the President's Emergency Fund for AIDS Relief (PEPFAR), DFID, and the International Development Association (IDA) at the World Bank. We specifically investigated the quality of development assistance for health (DAH) program evaluations published between 2009 and 2014 for their *relevance*, *validity*, and *reliability*. We intentionally chose these three criteria for assessing evaluation quality because they readily encompass the range of purposes and characteristics sought by different evaluators even when they disagree over relative importance (Rossi et al. 1999; Campbell 1969; Cronbach 1982; Patton 1997).

We defined evaluations as *relevant* if the authors addressed questions related to the means or ends of a program or intervention, and used appropriate data to answer those questions. We defined evaluations as *valid* if analyses were methodologically sound and conclusions were derived logically and consistently from the findings. We defined evaluations as *reliable* if the method and analysis would be likely to yield similar conclusions if the evaluation were repeated in the same or similar context.

Unlike systematic reviews that pre-specify an acceptable set of methodologies, we assessed whether an evaluation produced relevant, valid, and reliable findings considering whether its approach met accepted social science methodological standards. For example, evaluations may do a better or worse job of addressing bias in the way they collect information regardless of whether they are qualitative or quantitative and whether they focus on impact or other performance criteria.

This approach is different than if we had assessed each evaluation based on the standards applied by each agency. For example, DFID applies a set of standards to the evaluations produced by its Evaluation Unit that differs from the standards the Independent Commission for Aid Impact (ICAI) applies to its reviews. Similarly, many of the World Bank's Independent Evaluation Group (IEG) evaluations aim to assess the accuracy of project completion reports and do not apply the same standards that are used for other kinds of evaluations.

Instead, we sought to apply standards that are general enough and important enough to reflect the quality of the information and logic by which the evaluation reached its conclusions. After assessing the quality of a sample of evaluations on this basis, we looked for characteristics that might be associated with better quality. In the final section of the paper, we then offer 10 recommendations for improving the quality of evaluations (which are also summarized in Box 1).

Box 1: Summary of 10 recommendations for improving evaluation quality

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

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 the report in terms of potential bias.

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

Plan and design the evaluation before program implementation begins: Early planning has a robust association with higher evaluation quality.

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.

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

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.

Improve data collection methods: Interviews with key informants, focus groups, and beneficiaries are rich sources of information for evaluators but many evaluations, particularly those using qualitative methods, rely on convenience samples rather than the kinds of purposeful or random sampling that provide more robust confidence in findings.

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. Quantitative evaluations make little sense without information about context, implementation and the meaning people ascribe to programs. On the other hand, qualitative evaluations lack credibility without information (often quantitative) that corroborates or qualifies the information from interviews and focus group discussions.

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.

Methodology

Sample

To assess the quality of evaluations, we selected a sample of global health program 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 programs because health comprises one of the largest components of ODA (about 14 percent in 2015)[1] and because the global health field is widely perceived as a pioneer in conducting evaluations (Cameron et al. 2016). We also chose to focus on evaluations of large-scale program implementation and to exclude evaluations testing smaller-scale novel interventions. While this subset is not representative of all funders or interventions, it represents the largest health-focused development assistance programs.

We further refined the universe of evaluations by focusing on evaluations published between 2009 and 2014 and that (1) were publicly available, (2) assessed some feature of impact or performance, and (3) whose subject was a large-scale health program implemented by a major funding agency, rather than a program implemented to conduct a study. To identify evaluations, we began by conducting a search of organization websites, Google, Google Scholar, and the 3ie Impact Evaluation Database. We also contacted the heads of evaluation departments at each organization to confirm that our list captured all relevant evaluations. Our search yielded evaluations of a broad variety of programs that funded direct services, in-kind medications, budget support, medical training, logistics support, infrastructure, and policy reforms. The universe of evaluations was finalized on September 1, 2015. Any evaluations published subsequently or which were otherwise brought to our attention after that time were not included.

A final aspect of delimiting the universe was to focus on five large funders of DAH: USAID, the Global Fund, PEPFAR, DFID, and IDA at the World Bank. In 2015, these funders committed US$18.2 billion out of the US$25.8 billion committed for all health aid to low- and middle-income countries (OECD DAC 2017). We initially sought to include the Bill and Melinda Gates Foundation (BMGF) in the study, but we excluded them when conversations with BMGF evaluations staff indicated that most BMGF funding went toward smaller-scale studies to test novel approaches, rather than to large-scale program implementation.

To generate a representative sample of evaluations, we classified them as either impact or performance evaluations. We defined "impact evaluations" as evaluations whose primary purpose is to measure the impact that can be attributed to a particular intervention. 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. From the list of 299 evaluations, we randomly chose five impact evaluations and five performance evaluations from each funder. We found that some funders (e.g., the Global Fund and

[1] Data accessed on June 7, 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).

PEPFAR)[2] had fewer than five evaluations, and in these cases, we assessed all available evaluations in each category.

Our final sample comprised 37 evaluations from five funders. This sample can be considered representative of evaluations of large-scale health programs implemented by funders who account for a substantial share of DAH. Nevertheless, variations across and within agencies in terms of the scope and methods applied in each evaluation may limit the extent to which findings can be generalized from the sample. Appendix 4 lists all the evaluations in the sample.

Assessment instrument

Seven evaluators trained in evaluative methods reviewed the sample of evaluations. We developed the assessment instrument based on social science standards for research (Maxwell 2004; King et al. 1994) and on the Cochrane rubric for assessing risk of bias (Higgins and Green 2011). Like 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 reviewers assess a subset of 11 evaluations and by revising the instrument based on our review to ensure that the questions and concepts were clear.

Once we finalized the instrument, two reviewers independently completed the assessment form included in Appendix 1 for each evaluation in the sample. Then, the reviewer pair met to discuss and resolve any discrepancies in their assessments. During this meeting, reviewers 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 program or intervention, and used appropriate 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 calculated four aggregate scores corresponding to these features: relevance of objectives, relevance of data, sampling validity and reliability, and analytical validity and reliability. Other outcomes of interest included ethics and potential conflicts of interest. For each outcome, we ranked evaluations as having a low, medium, or high score by aggregating the reviewer pair's responses to multiple questions from the assessment form, as indicated in Appendix 2. The reviewer scores database and computer code used to analyze data are available at https://www.cgdev.org/media/assessing-quality-global-health-evaluations.

[2] Note that most agencies have more evaluations that do not meet all of our criteria for inclusion, such as internal evaluations which are not published or which were published after our universe was identified. In addition, the included PEPFAR evaluations were contracted under the auspices of USAID.

Analysis

We first analyzed descriptive statistics, focusing on the number of publicly available evaluations from each organization relative to their commitments to DAH. We then analyzed the four aggregate scores: relevance of data, relevance of objectives, sampling validity and reliability, and analytical validity and reliability. The sample was disaggregated 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 highest score. 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 two. The rationale for this approach was that a low or medium score on one component could undermine that entire aspect of the analysis.

To assess the reliability of our own assessments, we calculated the percent agreement for each outcome and its components (Di Eugenio and Glass 2004; Viera et al. 2005). We then 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 initially planned to assess the relationship between evaluation and program costs and quality but were unable to obtain precise data on the costs of the evaluations or the programs they evaluated. We present our pre-analysis plan in Appendix 3.

Findings

Table 1 compares evaluations in our universe and sample, disaggregated by funder and evaluation type. The US Government provides more DAH through USAID and PEPFAR than any other funder, averaging US$7.4 billion per year between 2010 and 2012.[3] USAID also published the most evaluations within the 2009 to 2014 period, with 254 evaluations, while PEPFAR had 16 evaluations. The Global Fund provided US$2.7 billion in DAH on average and published 4 evaluations. DFID provided US$1.5 billion and published 14 evaluations. The World Bank provided US$1.1 billion and published 24 evaluations. The scope and nature of these evaluations differed widely across and even within agencies. For example, one Global Fund report evaluated all interventions used to address AIDS, tuberculosis, and malaria in 18 countries over five years, while one USAID report evaluated a pilot intervention using community-based distribution of Misoprostol to prevent post-partum hemorrhage (see Appendix 4).

[3] We looked at average commitments between 2010 and 2012 because it corresponds to the mid-point of the publication period of the evaluations included in our universe.

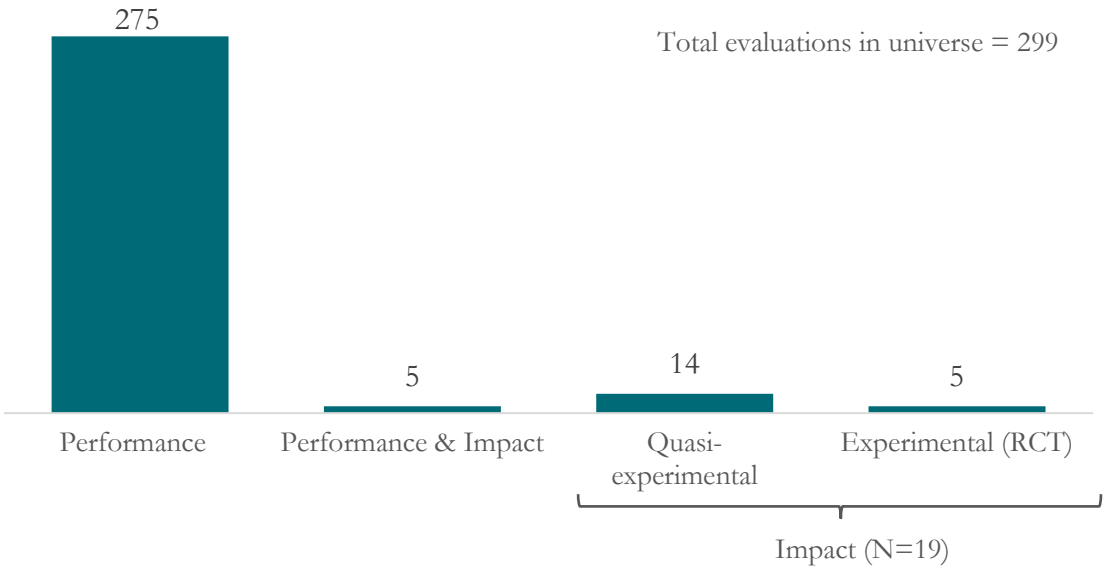
Table 1: Universe and sample of evaluations: 2009-2014

Funder	Average annual DAH 2010-12 (billions)	Universe				Sample			
		Total N	Impact N (%)	Perfor-mance N (%)	Both N (%)	Total N	Impact N (%)	Perfor-mance N (%)	Both 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: OECD. Creditor Reporting System. Accessed Jan. 30, 2017.
Note: In subsequent analyses, we treated the three evaluations in our sample labelled "both" as "impact" evaluations; reviewers were asked to focus on the "impact" evaluation sections of those studies.

Figure 1 below provides a breakdown of the evaluations in our universe by type and methodology. Most evaluations (92 percent) in our universe were performance evaluations, reflecting the weight of USAID studies in our universe and comparable to the share found in all USAID evaluations in an earlier study (Hageboeck et al. 2013). Out of 299 evaluations, 275 were performance evaluations, 19 were impact evaluations, and 5 were a combination. See the note below Figure 1, which explains why the breakdown in Figure 1 differs slightly from that in Table 1. Among the 19 impact evaluations in our universe, 5 used experimental methods and 14 used quasi-experimental approaches. Most quasi-experimental evaluations (8 out of 14) used a difference-in-differences approach. USAID had the smallest share of impact evaluations (4 percent), while the World Bank had the largest share (30 percent).

Figure 1: All evaluations in universe by type and methodology



Source: Authors' tabulations.
Notes: RCT = randomized controlled trial. The categorization presented in this figure differs slightly from the breakdown in 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 categorized as "performance & impact" were included in our sample and treated as "impact" evaluations for the analysis.

The second panel of Table 1 presents the 37 evaluations in our sample by funder and type. Our sample included a total of 21 performance evaluations, 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 that a handful of evaluations were initially misclassified. Furthermore, for the three evaluations in our sample which 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.

Table 2 presents the aggregate scores for relevance, reliability, and validity of the 37 evaluations in our sample. Most evaluations had *relevant objectives*. Three-quarters of the evaluations (28) addressed questions that reflected the main objectives of the program being evaluated. Only one evaluation was scored as addressing questions unrelated to the program objectives. Second, we assessed the *relevance of data* by considering whether the information that was collected and used reflected the population being served or services being provided. We found that 43 percent of the evaluations (16) failed to use any data on program recipients, goods and services, or health outcomes. Furthermore, more than 40 percent (15) used some data in these categories and were judged to at least "partially" use the data analytically to address the main research questions. Of the six evaluations (16 percent) which met the highest level of data relevance, all were impact evaluations.

In terms of the collection and use of information, 22 percent of the evaluations (8) met the highest standard for *sampling validity and reliability*, 11 percent (4) had a score in the middle range, and 68 percent (25) had the lowest score. To reach the highest score, evaluations had to meet several criteria including: use purposeful sampling, provide adequate methodological justification, and seek data from a sufficiently heterogeneous set of sources. Evaluations with the lowest score used convenience sampling without justification or lacked information on the data collection methods. About one-fifth of the evaluations (8) scored high on *sampling validity and reliability*—38 percent of the impact evaluations and 10 percent of the performance evaluations.

Finally, we assessed *analytical validity and reliability* primarily in terms of whether evaluations addressed potential confounding factors and adequately explained their methods of analysis. 16 percent of the evaluations (6) had high analytical validity and reliability. Another 35 percent (13) scored in the middle range and 49 percent (18) received a low score. We found that about a third (31 percent) of impact evaluations and 5 percent of performance evaluations had high analytical validity. Most performance evaluations (71 percent) had low analytical validity, while half of impact evaluations had medium validity (50 percent).

Table 2: Summary scores for evaluations in sample (N= 37)

Scores	Evaluation type	Low N (%)		Medium N (%)		High N (%)	
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%

Notes: Percentages may add to more than 100 percent due to rounding

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

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

We also investigated the association between prior planning of evaluations, the use of baseline data, and the use of monitoring data with each of the four aggregate scores using linear regression analysis (see Table 4). Evaluations indicated they were planned prior to program implementation in 24 percent of the cases (9 evaluations); only one of these was a performance evaluation. Prior planning of evaluations was positively associated with increased use of relevant data (2.23, 95 percent CI: 1.30 – 3.82) and better analytical validity and reliability (2.67, 95 percent CI: 1.57-4.53), relative to evaluations that were not planned prior to program implementation. Use of baseline data was mentioned in 30 percent of the cases (1 performance evaluation and 10 impact evaluations). Sixty-five percent of the evaluations—16 performance evaluations and 8 impact evaluations—mentioned use of monitoring data. The use of baseline or monitoring data was not statistically associated with better evaluation quality.

Table 4: Associations between evaluation characteristics and quality measures

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

Note: We conducted linear regression analyses controlling for evaluation type; results use robust standard errors.
*** = <0.01

Secondary outcomes of interest included consideration of ethics and data transparency. In 88 percent 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 46 percent (17) of the evaluations mentioned ethics considerations in data collection and that only 22 percent (8) received institutional review board approval.

Discussion

Global health program evaluations are key to learning about the best approaches for scaling up and delivering health programs. 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 programs. We found that most evaluations were performance evaluations rather than impact evaluations; among impact evaluations, experimental evaluations (randomized controlled trials or RCTs) constitute a minority. In addition, fewer than half of impact evaluations and fewer than 10 percent of performance evaluations met social science standards for relevance, validity, or reliability. Despite concerns that impact evaluations are less relevant than other kinds of evaluation (Cronbach 1982), 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 program evaluations to better promote learning and accountability.

Prior planning

One of the most basic ways to improve evaluation quality involves better and early planning. Nine evaluations (of which only one was a performance evaluation) referenced evaluation planning prior to program implementation. The remaining evaluations lacked references to pre-program 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 percent (13) of them interviewed program beneficiaries. Of those that interviewed beneficiaries, 40 percent used convenience sampling, which can easily lead to biased findings. For 18 performance evaluations in our sample, reviewers judged that interviewing program non-recipients would have enhanced the validity of the findings, but only four of them did so. Performance evaluations could be improved through more deliberate and preferably representative sampling of both program recipients and non-recipients.

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 program evaluation community (Brass et al. 2008). In our review of program evaluations by five large funders of DAH, we found that the spectrum of evaluation types is heavily weighted toward performance evaluations. In our universe of 299 health program evaluations, less than 2 percent (5) were RCTs, while almost 5 percent (14) used quasi-experimental methods. In contrast, 92 percent were performance evaluations, most of which used qualitative methods. These performance evaluations often had lower scores on relevance of data, sampling validity and reliability, and analytical validity and reliability.

Impact evaluations are achieving better methodological rigor even though deficiencies persist. By contrast, methodological rigor 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 rigor 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 conducted or commissioned by funders to assess their own projects, which can encourage positive bias (White and Bamberger 2008). Operational staff conducted some of these evaluations, but most came from offices that are formally separated from operational departments to mitigate this problem. For example, the World Bank's IEG reports directly to the institution's Board of Directors, not the President; Britain created ICAI, 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 percent of the cases (6 of the 37 evaluations). However, 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 people, which raises potential ethical issues. For example, interviews with individuals who have socially stigmatized 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 to inter-rater differences. 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 generalizable 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 program evaluations to encompass different funders. Our method of assigning multiple reviewers and testing instruments generated consistent findings, which increases the credibility of the results.

Despite our efforts to generate a representative sample of evaluations, the inherent diversity of aims, methods, and institutional standards limits the extent to which the findings can be generalized. Furthermore, the large amount of time it took to assess evaluations (about 12

full-time equivalent weeks of preparation, reading, and discussion) led to a relatively small sample, only 37 evaluations. This limited our ability to identify consistent patterns, compare funders, and differentiate among different evaluation methods. The small sample size also made it impossible to test for trends in evaluation quality. Finally, the sample was based on published evaluations from 2009 to 2014, and therefore does not reflect changes which may have occurred since then due to new evaluation policies, institutions, or leadership at the different agencies.

Ten recommendations for agencies to improve evaluations of development programs

Drawing on our findings and analysis, we have developed a set of practical recommendations, which may serve as a checklist for agency staff overseeing and managing evaluations.

Classify the evaluation purpose: Be clear at meta-level data (e.g., title, abstract, coding/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 other questions about a program's performance, such as effective implementation, management, or satisfaction.

In our assessment, we differentiated between "impact evaluations" and "performance evaluations"; in practice however, other terms are also used. We found that three of the evaluations in our universe we initially classified as impact evaluations were actually performance evaluations; another three evaluations were described as performance evaluations but contained data and analysis that qualified them as impact evaluations.

In addition to including information in the title and abstract, a tagging system that distinguishes these types of evaluations could be used in online evaluation databases (e.g., USAID's Development Experience Clearinghouse). In this way, people searching for evidence of impact (a more general audience) can easily identify the evaluations they are looking for. By contrast, performance evaluations tend to be of greater interest to specific audiences within agencies or governments. Distinctions between types of evaluations should follow the definitions laid out in each agencies' evaluation policy (e.g., USAID, DFID, etc.).

Discuss evaluator independence: Evaluations should provide clear information about the evaluators' institutional affiliation and financial conflicts of interest so that readers can assess the report in terms of potential bias. Furthermore, the involvement, if any, of implementers in sampling, data collection, or report write-up should be explicitly acknowledged if it occurs. We judged that evaluators had significant independence in 22 of the 37 evaluations (60 percent) in our sample, while evaluators in six of the evaluations (16 percent) did not. Nevertheless, we were unable to make a judgment about the degree of independence in nine of the evaluations (24 percent) due to incomplete information.

Disclose costs and duration of programs and evaluations: Two of the most basic pieces of information about a development program are its cost and duration. The evaluations that we reviewed frequently lacked this information (Table 5). Program cost information was only available for 20 evaluations in our sample (54 percent). Details about program duration were available for 23 of the evaluations (62 percent). Furthermore, managing the evaluation process should be guided by some sense of the benefits (the value of the information) relative to the costs of obtaining it. Only one evaluation included information about the financial resources required to conduct the evaluation. Even the duration of the evaluation itself was missing or unclear in 25 of the 37 evaluations (67 percent).

Table 5: Information on cost and duration of programs and evaluations

Question	No		Partially		Yes	
Is there information on the cost of the program that was evaluated?	30%	11	16%	6	54%	20
Is there information on the duration of the program that was evaluated?	8%	3	30%	11	62%	23
Is there information on the cost of the evaluation?	92%	34	5%	2	3%	1
Is there information on the duration of the evaluation?	19%	7	49%	18	32%	12

Note: The total number of evaluations in the sample was 37.
Source: Tabulation by authors.

Plan and design the evaluation before program implementation begins: The relationship between the quality of evaluations and planning the evaluation prior to the start of the project is both plausible and confirmed statistically by our review. Evaluators themselves mentioned the lack of baseline data (27 percent), lack of monitoring data (14 percent), insufficient time (30 percent), and insufficient budget (11 percent) as evaluation limitations. Baseline data is particularly important. Only 11 evaluations (of which one was a performance evaluation) mentioned use of baseline data. Yet performance evaluations also benefit when, for example, practices of managers or staff at the end of a project can be compared to practices at the beginning. Taking the time at the beginning of a project to establish initial questions, lay out an evaluation timeline including plans for data collection, and budget funds for a proper evaluation is likely to lead to more useful findings without necessarily increasing costs.

State the evaluation question(s) clearly: Perhaps the most important starting point for any evaluation is the question(s) being asked. A clear and well-defined question does more than help 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. We found that only 18 evaluations (49 percent) clearly listed evaluation questions.

Explain the theoretical framework underpinning the evaluation: Evaluations are easier to assess if evaluators are explicit about the theoretical framework that underlies their analysis. In many cases, this simply requires reference to an existing literature. In other cases, evaluators have substantially adapted or proposed new theories. We found that that 24 evaluations (65 percent) did not apply a theoretical framework; 5 evaluations (13 percent) applied a framework based on existing literature and the remaining 8 evaluations (22 percent) applied a framework developed specifically for that evaluation.

Explain sampling and data collection methods: Evaluations should be clear enough about sampling and data collection so that subsequent researchers could replicate them if desired. We found that only 7 of the 37 evaluations (19 percent) adequately explained the data collection process. The remaining did not include any such details (15 evaluations) or only provided a partial explanation (15 evaluations). The purpose of documenting sampling procedures in social science research is rarely intended to actually replicate the same evaluation. Rather, the process of documentation allows readers to assess the quality of the information collected; provides guidance and ideas to future evaluators; and, in some instances, makes it possible to conduct follow-up evaluations.[4]

Improve data collection: Interviews with key informants, focus groups, and beneficiaries are rich sources of information for evaluators but many evaluations, particularly those using qualitative methods, rely on convenience samples rather than the kinds of purposeful or random sampling that provide more robust confidence in findings. Overall, we found that 11 evaluations used convenience sampling, 16 of them used purposeful sampling, 13 used random sampling, and we were unable to determine the sampling methodology for 9 evaluations (8 performance and 1 impact).

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. Quantitative evaluations make little sense without information about context, implementation and the meaning people ascribe to programs. On the other hand, qualitative evaluations lack credibility without information (often quantitative) that corroborates or qualifies the information from interviews and focus group discussions. Of the 37 evaluations we reviewed, only 12 (32 percent) used a combination of interviews, focus groups, and surveys. Two evaluations (5 percent) reported that they only used interviews or focus groups and five (14 percent) reported only using surveys.

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. Data certainly needs to be

[4] An interesting example in which evaluators actually replicated an earlier impact evaluation is referenced in Independent Commission for Aid Impact (ICAI) 2013, "DFID's Livelihoods Work in Western Odisha."

published with care to address legitimate concerns regarding confidentiality and privacy.[5] Only two of the 37 evaluations in our sample indicated that data was publicly available, although some data may have been published subsequently.

The possibility that evaluations might harm human subjects is also a legitimate concern for which certain procedures and guidelines have been developed. Nevertheless, only 8 of the 37 evaluations (22 percent) in our sample mentioned they were reviewed independently for ethical concerns, while 17 of them (46 percent) mentioned that they had considered and addressed ethical concerns in data collection.

Conclusion: Achieving learning and accountability with global health program evaluations

Evaluations are critical to learning about what works in global development and for holding funders and implementers accountable. We reviewed evaluations from five large DAH funders, which together commit about US$12-18 billion to health programs 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 the share of evaluations that meet social science standards to assure the relevance, validity, and reliability of findings. A key recommendation from our analysis is to plan more evaluations before program implementation begins. We also encourage the global development community to have an intensive discussion about the methodological quality of performance evaluations, which make up 92 percent of publicly available evaluations from five large DAH funders. We also encourage them to focus on the need to better invest in collecting and using robust and varied types of data. It is possible that better planning while using the same amount of evaluation funds could yield more relevant, valid, and reliable information to guide future programs. Finally, funders and program evaluators should more explicitly address conflicts of interest and follow appropriate standards for protecting human subjects. Taking these steps to improve evaluations is critical to ensuring that global health interventions, and development programs more generally, effectively and efficiently contribute to improved lives and increased wellbeing around the world.

[5] See, for example, Sturdy, Jennifer, Stephanie Burch, Heather Hanson, and Jack Molyneaux. 2017. "Opening up Evaluation Microdata: Balancing Risks and Benefits of Research Transparency." BITSS. March 3. https://osf.io/preprints/bitss/s67my.

References

3ie (International Initiative for Impact Evaluation). 2015. *3ie Annual Report 2015: Evidence, Influence, Impact*. New Delhi: 3ie.

Andrews, Matt, Lant Pritchett, and Michael Woolcock. 2012. "Escaping Capability Traps through Problem-Driven Iterative Adaptation." Working Paper 299. Washington, DC: Center for Global Development.

BMGF (Bill and Melinda Gates Foundation). 2017. Evaluation Policy. Seattle, WA: Bill and Melinda Gates Foundation.

Burnside, Craig and David Dollar. 2000. "Aid, Policies, and Growth." *American Economic Review*, September, 90(4):847– 68.

Cameron, Drew B., Anjini Mishra, and Annette N. Brown. 2016. "The Growth of Impact Evaluation for International Development: How Much Have We Learned?" *Journal of Development Effectiveness* pp. 1-21.

Campbell, Donald. 1969. "Reforms as Experiments." *American Psychologist* vol. 24: 409–429.

Cohen, J. and W Easterly. 2010. *What Works in Development? Thinking Big and Thinking Small.* Brookings Institution Press.

Cronbach, Lee J., and Karen Shapiro. 1982. *Designing evaluations of educational and social programs*. Jossey-Bass.

DFID (UK Department for International Development). 2013. DFID Evaluation Policy 2013. London: DFID.

Di Eugenio, Barbara and Michael Glass. 2004. "The kappa statistic: A second look." *Computational Linguistics* 30.1: 95-101.

Doemeland, Doerte and James Trevino. 2014. "Which World Bank Reports Are Widely Read?" Policy Research Working Paper 6851. Washington, DC: World Bank.

Dollar, David, and Aart Kraay. 2003. "Institutions, trade, and growth." *Journal of Monetary Economics,* 50.1: 133-162.

Easterly, William, Ross Levine, and David Roodman. 2004. "Aid, Policies, and Growth: Comment." *American Economic Review*, June, 94(3):774–80.

Easterly, William. 2001. *The Elusive Quest for Growth: Economists' Adventures and Misadventures in the Tropics*. Boston, MA: MIT Press.

EGWG (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.

GAO (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.

Gertler PJ, S. Martinez, P. Premand, L. B. Rawlings, and C. M. Vermeersch. 2016. *Impact Evaluation in Practice*. Washington, DC: World Bank Publications.

Glassman, Amanda, and Miriam Temin. 2016. *Millions Saved: New cases of Proven Success in Global Health*. Washington, DC: Brookings Institution Press.

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

Higgins, J.P., and S. Green. 2011. *Cochrane Handbook for Systematic Reviews of Interventions*. Vol 4. John Wiley & Sons.

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

IEG (Independent Evaluation Group). 2012. "World Bank Group Impact Evaluations: Relevance and Effectiveness." Washington DC: World Bank.

Kennedy-Chouane, Megan and Hans Lundgren. 2013. Evaluating Development Activities - 12 Lessons from the OECD DAC. Paris, France: OECD.

King, G., R. O. Keohane, S. Verba. 1994. *Designing Social Inquiry: Scientific Inference in Qualitative Research*. Princeton, NJ: Princeton University ress.
Levine, Ruth and William Savedoff. 2015. "Aid at the Frontier: Building Knowledge Collectively." *Journal of Development Effectiveness*, 7:3, pp. 275-289.
Maxwell JA. 2004. "Using qualitative methods for causal explanation." *Field Methods*, 16(3):243-264.
OECD DAC (OECD Development Assistance Committee). 2005. *The Paris Declaration on Aid Effectiveness.* Paris.
OECD DAC. 2010. "Quality Standards for Development Evaluation." DAC Guidelines and Reference Series. Paris, France: OECD.
OECD DAC. 2016. *OECD Development Co-operation Peer Reviews: United States 2016*. Paris, France: OECD.
OECD DAC. 2017. Creditor Reporting System. Accessed June 7, 2017.
Patton, Michael Quinn. 1997. *Utilization-Focused Evaluations: The New Century Text.* Thousand Oaks, CA: Sage Publications.
Rossi, Peter Henry, Mark W. Lipsey, and Howard E. Freeman. 1999. *Evaluation: A Systematic Approach*. Sage publications.
Scriven, Michael. 1991. *Evaluation Thesaurus* 4th ed. Newbury Park, CA: Sage Publications.
USAID (United States Agency for International Development). 2016. USAID Evaluation Policy. Washington, DC: USAID.
USAID. 2016. "Five Years of Evaluation Practice." Washington DC: USAID.
Viera, Anthony J., and Joanne M. Garrett. 2005. "Understanding interobserver agreement: the kappa statistic." *Fam Med* 37.5: 360-363.
White, H. and M. Bamberger, 2008. "Introduction: impact evaluation in official development agencies." *IDS Bulletin*, 39(1), p.1.

Appendix 1: Evaluation assessment form

Available here: https://docs.google.com/forms/d/e/1FAIpQLSf2zddmQ5ToJD2v7sWqq5H_iUU8sU-lR4NcVb6Vl9S_ALXhzg/viewform

Appendix 2: Methodology for constructing scores

3 = all 3s; **2** = one 2 **1 =** two or more scores of 2 or less, one or more scores of 1
Note: The rationale for the scoring is that low/medium scores in any category may bias the entire analysis

Relevance of objectives score

3	Do the evaluation questions and objectives/ purpose reflect the main objectives of the program being evaluated?	Y
2	Do the evaluation questions and objectives/ purpose reflect the main objectives of the program being evaluated?	Partially
1	Do the evaluation questions and objectives/ purpose reflect the main objectives of the program being evaluated?	N

Relevance of data score

3	Was data collected on the following subjects if relevant? Program recipients Program non-recipients Goods and services	More than 1 = Y
	Does the evaluation use secondary data? & Is the data relevant to the program being evaluated?	If Y Then Y
	Is the analytical approach appropriate for answering the research question/objective?	Y
	Were any of the following types of data used in the analysis? Data on whether people used goods and services provided/ on training outcomes Data on knowledge or behavior change Data on health outcomes	More than 1 = Y
2	Was data collected on the following subjects if relevant? Program recipients Program non-recipients Goods and services	At least 1 = Y
	Is the analytical approach appropriate for answering the research question/objective?	Partially
	Were any of the following types of data used in the analysis? Data on whether people used goods and services provided/ on training outcomes Data on knowledge or behavior change Data on health outcomes	At least 1 = Y

1	Was data collected on the following subjects if relevant? Program recipients Program non-recipients Goods and services	None
	Does the evaluation use secondary data? & Is the data relevant to the program being evaluated?	Y N
	Is the analytical approach appropriate for answering the research question/objective?	N
	Were any of the following types of data used in the analysis? Data on whether people used goods and services provided/ on training outcomes Data on knowledge or behavior change Data on health outcomes	None

Sampling validity and reliability score

3	Random sampling	Y
	Purposeful sampling & Do the authors provide adequate methodological justification for their sampling approach? & If the authors used purposeful sampling, did they seek heterogeneous populations?	Y Y Y
2	Convenience sampling & Do the authors provide adequate methodological justification for their sampling approach?	Y Y
	Purposeful sampling & Do the authors provide adequate methodological justification for their sampling approach? If the authors used purposeful sampling, did they seek heterogeneous populations?	Y If either or both = somewhat
1	Convenience sampling & Do the authors provide adequate methodological justification for their sampling approach?	Y N / somewhat
	Purposeful sampling & Do the authors provide adequate methodological justification for their sampling approach? If the authors used purposeful sampling, did they seek heterogeneous populations?	Y If either or both = N
	Which sampling methods were used?	Don't know

Analytical validity and reliability score

3	If randomized, is the randomization method fully described?	Y
	If randomized, is the randomization method fully described? & Are all appropriate covariates that may influence findings considered in the analysis?	N Y
	If not randomized, are all appropriate covariates that may influence findings considered in the analysis?	Y
2	If randomized, is the randomization method fully described? & Are all appropriate covariates that may influence findings considered in the analysis?	N Partially
	If not randomized, are all appropriate covariates that may influence findings considered in the analysis?	Partially
1	If randomized, is the randomization method fully described? & Are all appropriate covariates that may influence findings considered in the analysis?	N N
	If not randomized, are all appropriate covariates that may influence findings considered in the analysis?	N

Ethics score

First determine if it is human subjects research		
Y	Did the evaluation use data on human subjects?	Y
Then determine score		
3	If monitoring data was used, was it de-identified? If the secondary data was used, was it de-identified?	Y Y
	Was primary data collected specifically for the evaluation? & Do the authors indicate whether they got ethics approval or exemption?	Y Y
1	If monitoring data was used, was it de-identified?	N
	If the secondary data was used, was it de-identified?	N

Reporting score

Score is the average of the following answers	
Are evaluation questions listed?	(1/2/3)
Are evaluation objectives listed?	(1/2/3)
Are terms in evaluation questions operationally defined?	(1/2/3)

Are sampling methods described in enough detail to replicate?	(1/2/3)
Does the evaluation involve random assignment to treatment? & Is the randomization method described?	(Y/N) (1/2/3)
Was primary data collected specifically for the evaluation? & Data collection instruments are identified & Data collection instruments are available?	(Y) (1/2/3)*
Does the evaluation use monitoring data from within the program? And… Data collection instruments are identified & Data collection instruments are available?	(Y/N) (1/2/3)
Does the evaluation use secondary data from sources external to the health program? And… The population of participants is fully described The data source is publicly available	(Y/N) (1/2/3)
Are the analytical methods described?	(1/2/3)
Are evaluation results presented for all questions?	(1/2/3)
Are estimates of error reported?	(1/2/3)
Are evaluation limitations discussed?	(1/2/3)

Appendix 3: Pre-analysis plan

Analysis Plan		
Number of publicly available evaluations		
Amount of development assistance for health in each year Relative to number of publicly available evaluations in each year		
Internal Validity		
3 = all 3s; **2** = one 2 **1 =** two or more scores of 2 or less, one or more scores of 1 *Note:* The rationale for the scoring is that low/medium scores in any category may bias the entire analysis		
Internal validity: Sampling		
3	Representative sampling	3
	Purposeful sampling & Do the authors provide adequate methodological justification for their sampling approach? If the authors used purposeful sampling, did they seek heterogeneous populations?	3 3
	Is there potential bias introduced by the funder/implementer role in sampling?	Unlikely
2	Convenience sampling & Do the authors provide adequate methodological justification for their sampling approach?	3
	Purposeful sampling & Do the authors provide adequate methodological justification for their sampling approach? If the authors used purposeful sampling, did they seek heterogeneous populations?	If either or both = 2
	Is there potential bias introduced by the funder/implementer role in sampling?	Unlikely
1	Is there potential bias introduced by the funder/implementer role in sampling?	Likely
	OR	
	Convenience sampling & Do the authors provide adequate methodological justification for their sampling approach?	(1/2)
	Purposeful sampling & Do the authors provide adequate methodological justification for their sampling approach? If the authors used purposeful sampling, did they seek heterogeneous populations?	If either or both = 1

Internal validity: Data		
3	Are the questions on the data collection instruments written in a manner that could capture heterogenous responses?	3
2	Are the questions on the data collection instruments written in a manner that could capture heterogenous responses?	2
1	Are the questions on the data collection instruments written in a manner that could capture heterogenous responses?	1
Internal validity: Analysis		
3	If randomized, was treatment assignment fully random?	Y
	If randomized, was treatment assignment fully random? & Are all appropriate covariates that may influence findings considered in the analysis?	N 3
	Are all appropriate covariates that may influence findings considered in the analysis?	3
2	If randomized, was treatment assignment fully random? & Are all appropriate covariates that may influence findings considered in the analysis?	N 2
	Are all appropriate covariates that may influence findings considered in the analysis?	2
1	If randomized, was treatment assignment fully random? & Are all appropriate covariates that may influence findings considered in the analysis?	N 1
	Are all appropriate covariates that may influence findings considered in the analysis?	1
Internal validity: Reporting		
3	Are the results presented in an objective manner?	3
	Are the conclusions presented in an objective manner?	3
	Is the conclusion consistent with the findings?	3
2	Are the results presented in an objective manner?	2
	Are the conclusions presented in an objective manner?	2
	Is the conclusion consistent with the findings?	2

1	Are the results presented in an objective manner?	1	
	Are the conclusions presented in an objective manner?	1	
	Is the conclusion consistent with the findings?	1	
	Internal validity consistency check		
	These scores will be used to cross-check Internal Validity scores but will not be part of the analysis		
	What is the risk of bias in this report? Explain	(1/2/3)	Including sampling bias, selection bias, and reporting bias
	Overall quality of analytical approach	(1-5)	1) No systematic analytical approach (2) Some effort at systematic analysis (3) Weak systematic analysis (applied but resulting in weak interpretation) (4) Good systematic analysis applied (with convincing interpretation) (5) Excellent systematic analysis (applied with convincing interpretation that also considers context and potential sources of error).
	What is your assessment of data quality, if you feel you can assess it?	(1/2/3/DK)	

	Relevance		
3	Are the evaluation questions/objectives consistent with the main program objectives?	3	
	Was data collected on at least one of the following subjects if relevant? Program recipients Program non-recipients Goods and services	Y	
	If the evaluation uses secondary data, is the data relevant to the program being evaluated?	3	
	Is the analytical approach appropriate for answering the research question/objective?	3	
	Were any of the following types of data used in the analysis? Data on whether people used goods and services provided/ on training outcomes Data on knowledge or behavior change Data on health outcomes	Y	
2	Are the evaluation questions/objectives consistent with the main program objectives?	2	
	Was data collected on at least one of the following if relevant? Program recipients Program non-recipients Goods and services	Y	
	If the evaluation uses secondary data, is the data relevant to the program being evaluated?	2	

	Is the analytical approach appropriate for answering the research question/objective?	2
	Were any of the following types of data used in the analysis? Data on whether people used goods and services provided/ on training outcomes Data on knowledge or behavior change Data on health outcomes	Y
1	Are the evaluation questions/objectives consistent with the main program objectives?	1
	Was data collected on at least one of the following if relevant? Program recipients Program non-recipients Goods and services	N
	If the evaluation uses secondary data, is the data relevant to the program being evaluated?	1
	Is the analytical approach appropriate for answering the research question/objective?	1
	Were any of the following types of data used in the analysis? Data on whether people used goods and services provided/ on training outcomes Data on knowledge or behavior change Data on health outcomes	N

External Validity		
3	Representative sampling	3
	Purposeful sampling and… Do the authors provide adequate methodological justification for their sampling approach?	3
	If the authors used purposeful sampling, did they seek heterogeneous populations?	3
	Is the external validity of the evaluation discussed?	(3/2)
	Does the report include conclusions/ recommendations beyond the scope of the project being evaluated?	Y
2	Convenience sampling and… Do the authors provide adequate methodological justification for their sampling approach?	3
	Purposeful sampling and…	

	Do the authors provide adequate methodological justification for their sampling approach? If the authors used purposeful sampling, did they seek heterogeneous populations?	If either or both = 2
	Is the external validity of the evaluation discussed?	2
	Does the report include conclusions/ recommendations beyond the scope of the project being evaluated?	Y
1	Convenience sampling and… Do the authors provide adequate methodological justification for their sampling approach?	(1/2)
	Purposeful sampling and… Do the authors provide adequate methodological justification for their sampling approach? If the authors used purposeful sampling, did they seek heterogeneous populations?	If either or both = 1
	Is the external validity of the evaluation discussed?	1
	Does the report include conclusions/ recommendations beyond the scope of the project being evaluated?	N

Open Data	
Is the data publicly available?	Y/N

Ethics		
First determine if it is human subjects research		
Y	Did the evaluation use data on human subjects?	Y
Then determine score		
3	If monitoring data was used, was it de-identified?	Y
	If the secondary data was used, was it de-identified?	Y
	Was primary data collected specifically for the evaluation? & Do the authors indicate whether they got ethics approval or exemption?	Y Y
1	If monitoring data was used, was it de-identified?	N
	If the secondary data was used, was it de-identified?	N
	Was primary data collected specifically for the evaluation? And… Do the authors indicate whether they got ethics approval or exemption?	Y N

Reporting Score	
Score is the average of the following answers	
Are evaluation questions listed?	(1/2/3)
Are evaluation objectives listed?	(1/2/3)
Are terms in evaluation questions operationally defined?	(1/2/3)
Are sampling methods described in enough detail to replicate?	(1/2/3)
Does the evaluation involve random assignment to treatment? And…	(Y/N)
Is the randomization method described?	(1/2/3)
Was primary data collected specifically for the evaluation? And…	(Y/N)
Are the data collection instruments identified and available?	(1/2/3)
Are the data collections instruments provided?	(1/2/3)
Does the evaluation use monitoring data from within the program? And…	(Y/N)
Are the data collection instruments identified and available?	(1/2/3)
Are the data collections instruments provided?	(1/2/3)
Does the evaluation use secondary data from sources external to the health program? And…	(Y/N)
Is the secondary data population fully described?	(1/2/3)
Is the secondary data source available?	(1/2/3)
Are the analytical methods described?	(1/2/3)
Are evaluation results presented?	(1/2/3)
Are estimates of error reported?	(Y/N)
Are evaluation limitations discussed?	(1/2/3)

Predictor Variables
Would you categorize the evaluators as academics, contractors, internal employees, or other?
Year of evaluation
Ratio of evaluation/program cost What is the program cost? What is the evaluation cost? If there is not information on evaluation cost, write in any other information that might inform evaluation cost, such as person hours and team members.
Ratio of evaluation/program duration What is the evaluation duration? What is the program duration?
Do the evaluators mention any of the following as constraints on evaluation methodology? Time constraints

Presence of monitoring and evaluation data Access to monitoring and evaluation data Quality of monitoring and evaluation data Budget constraints Lack of access to project information Other (specify)
Do the evaluators indicate that they started planning the evaluation prior to or at the beginning of the health program?
Was baseline data collected as part of the evaluation?
Which of the following data collection methods were used, regardless of whether data is primary or secondary? Interviews Focus groups Surveys Routine monitoring data Facility records Observation Other (clarify)

Appendix 4: List of 37 evaluations included in our sample

Evaluation title	Year	Funder	Type	URL
DFID's Work through the United Nations Children's Fund (UNICEF)	2013	DFID	performance	http://icai.independent.gov.uk/wp-content/uploads/ICAI-report-DFIDs-work-with-UNICEF.pdf
Final Evaluation of the Three Diseases Fund	2012	DFID	performance	https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/204884/Evaluation-three-diseases-fund-and-management-response.pdf
Evaluation of a Decade of DFID and World Bank Supported HIV and AIDS Programmes in Vietnam from 2003 to 2012	2013	DFID	impact	https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/303560/Decade-DFID-World-Bank-Support-HIV-Aids-Prog-Vietnam-2003-2012.pdf
DFID's Water, Sanitation, and Hygiene Programming in Sudan	2013	DFID	performance	http://icai.independent.gov.uk/wp-content/uploads/ICAI-Report-DFIDs-Water-Sanitation-and-Hygiene-Programming-in-Sudan.pdf
COLALIFE Operational Trial Zambia: Endline Survey Report	2014	DFID	both	https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/347891/Evaluation-Cola-Life-Trial-Zambia.pdf
External Evaluation Report of Sierra Leone's Youth Reproductive Health Programme (2007 – 2012)	2013	DFID	impact	https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/206755/eval-resteless-dev-Sierra-Leone-youth-rep-health-prog-2007-2012.pdf
Southern African Regional Social and Behavior Change Communication Program	2013	DFID	impact	http://r4d.dfid.gov.uk/pdf/outputs/SAR_Evaluation/Mozambique_SBCC_2013_FV.pdf
The Department for International Development's Support to the Health Sector in Zimbabwe	2011	DFID	both	http://icai.independent.gov.uk/wp-content/uploads/DFIDs-Support-to-the-Health-Sector-in-Zimbabwe3.pdf
Evaluation of DFID's Support for Health and Education in India	2012	DFID	performance	http://icai.independent.gov.uk/wp-content/uploads/ICAI-Evaluation-of-DFIDs-Support-for-Health-and-Education-in-India-Final-Report.pdf
Impact Assessment of the Expanded Support Programme, Zimbabwe	2011	DFID	performance	https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/197475/ESP-Impact-Assessment-zimb-11.pdf
Five-Year Evaluation Study Area 3 Results: Scaling Up Against HIV, TB, and Malaria	2009	Global Fund	impact	https://www.theglobalfund.org/media/2978/terg_sa3_report_en.pdf
Independent Evaluation of the Affordable Medicines Facility - malaria (AMFm) Phase 1	2012	Global Fund	impact	https://www.theglobalfund.org/media/3011/terg_evaluation2013-2014thematicreviewamfm2012iephase1_report_en.pdf

Evaluation of Global Fund Investments in Country Monitoring and Evaluation Systems	2012	Global Fund	performance	https://www.theglobalfund.org/media/3008/terg_evaluation2013-2014thematicreviewinvestinginmnesystems_report_en.pdf
PEPFAR Public Health Evaluation - Care and Support: Phase 2 Uganda	2010	PEPFAR*	impact	http://www.cpc.unc.edu/measure/resources/publications/tr-10-74d
USAID/Office of HIV and AIDS: Project SEARCH End of Project Evaluation Supporting Evaluation and Research to Combat HIV	2012	PEPFAR*	performance	http://pdf.usaid.gov/pdf_docs/PA00JWRQ.pdf
USAID/South Africa Umbrella Grants Management Project, End of Project partner Evaluation: Senzakwenzeke Community Development	2012	PEPFAR*	performance	http://pdf.usaid.gov/pdf_docs/PDACX481.pdf
USAID/South Africa Umbrella Grants Management Project, End of Project partner Evaluation: Project Concern International	2013	PEPFAR*	performance	http://pdf.usaid.gov/pdf_docs/PDACX455.pdf
USAID/South Africa Umbrella Grants Management Project, End of Project Partner Evaluation: Noah (Nurturing Orphans of AIDS for Humanity)	2012	PEPFAR*	performance	http://pdf.usaid.gov/pdf_docs/PDACX461.pdf
A Concurrent Evaluation of Phase II of the NRHM BCC Campaign (India)	2009	USAID	performance	http://pdf.usaid.gov/pdf_docs/PDACQ561.pdf
Nurturing the Mother-Child Dyad in an Urban Setting: Final Evaluation of the Hati Kami Project in Jakarta, Indonesia	2014	USAID	both	https://dec.usaid.gov/dec/GetDoc.axd?ctID=ODVhZjk4NWQtM2YyMi00YjRmLTkxNjktZTcxMjM2NDBmY2Uy&rID=MzU2NTQ4&pID=NTYw&attchmnt=VHJ1ZQ==&uSesDM=False&rIdx=NDU5ODc3
Community-Based Distribution of Misoprostol for the Prevention of Postpartum Hemorrhage: Evaluation of a Pilot Intervention in Tangail District, Bangladesh	2010	USAID	performance	http://pdf.usaid.gov/pdf_docs/PA00JRX6.pdf
USAID/Ethiopia: Implementing Partners' Organizational Capacity Assessment Report	2011	USAID	performance	http://pdf.usaid.gov/pdf_docs/PDACU381.pdf
Expanded Impact Child Survival Program, Final Evaluation Report Gaza Province, Mozambique	2009	USAID	performance	http://pdf.usaid.gov/pdf_docs/pdacp466.pdf
Impact Evaluation of the Project 'Strengthening sustainable orphans and vulnerable children (OVC) care and support in Cote d'Ivoire in the Urban Context of Abidjan: Final Evaluation Report	2014	USAID	impact	http://pdf.usaid.gov/pdf_docs/PA00K6Z6.pdf

Impact Evaluation of the Mayer Hashi Program of Long-Acting and Permanent Methods of Contraception in Bangladesh	2014	USAID	impact	http://pdf.usaid.gov/pdf_docs/PA00K269.pdf
Evaluation of the Quality of Community based Integrated Management of Childhood Illness and Reproductive Health Programs in Madagascar	2013	USAID	performance	http://pdf.usaid.gov/pdf_docs/PDACY055.pdf
The Impact of Training Personnel to the Minimum Standards ISPO Category I & II: Tanzania Training Centre for Orthopaedic Technologists	2012	USAID	performance	http://pdf.usaid.gov/pdf_docs/pbaaa691.pdf
Bangladesh Smiling Sun Franchise Program Impact Evaluation Report	2012	USAID	impact	http://pdf.usaid.gov/pdf_docs/PDACU705.pdf
Project Performance Assessment Report: Jamaica Social Safety Net Project (LN 70760) and National Community Development Project (LN71480)	2010	World Bank	performance	https://ieg.worldbankgroup.org/Data/reports/jamaica_ssn_ppar.pdf
Impact Evaluation of Three Types of Early Childhood Development Interventions in Cambodia	2013	World Bank	impact	https://openknowledge.worldbank.org/bitstream/handle/10986/15900/WPS6540.pdf?sequence=1
A Randomized, Controlled Study of a Rural Sanitation Behavior Change Program in Madhya Pradesh, India	2013	World Bank	impact	http://documents.worldbank.org/curated/en/300331468269410084/pdf/WPS6702.pdf
Project Performance Assessment Report for the Romania Avian Influenza Control and Human Pandemic Preparedness and Response Project	2013	World Bank	performance	https://ieg.worldbankgroup.org/Data/reports/PPAR-78781-P100470-Romania_Avian_Influenza.pdf
Peru Basic Health & Nutrition Project (Loan 3701) and Mother & Child Insurance and Decentralization of Health Services, First Phase of Health Reform Project (Loan 4527) Project Performance Assessment Report	2009	World Bank	performance	https://ieg.worldbankgroup.org/Data/reports/peru_health_ppar.pdf
Project Performance Assessment Report: Lesotho Health Sector Reform and HIV and AIDS Capacity Building and Technical Assistance Projects	2010	World Bank	performance	https://ieg.worldbankgroup.org/Data/reports/PPAR_Lesotho_Health_Sector_Reform_and_HIV-AIDS_Cap_Bldg.pdf
Project Performance Assessment Report: Ecuador - Rural and Small towns Water Supply and Sanitation Project	2011	World Bank	performance	https://ieg.worldbankgroup.org/Data/reports/PPAR_Ecuador_Rural_small_towns_water_supply_Sanitation.pdf
Impact Evaluation of a Large-Scale Rural Sanitation Project in Indonesia	2013	World Bank	impact	https://openknowledge.worldbank.org/bitstream/handle/10986/13166/wps6360.pdf?sequence=1
Impact Evaluation of School Feeding Programs in Lao PDR	2011	World Bank	impact	https://openknowledge.worldbank.org/bitstream/handle/10986/3291/WPS5518.pdf?sequence=1

Note: For the analysis, we categorized the three evaluations in our sample labelled "both" as "impact" evaluations.
*PEPFAR evaluations were conducted under the auspices of USAID.