USAID
FROM THE AMERICAN PEOPLE

CREDIT: USAID/ETHIOPIA

ASSESSING THE QUALITY OF IMPACT EVALUATIONS AT USAID

DECEMBER 22, 2020

This publication was produced for review by the United States Agency for International Development. It was prepared for the E3 Analytics and Evaluation Project by Management Systems International, A Tetra Tech Company.

(THIS PAGE INTENTIONALLY LEFT BLANK)

ASSESSING THE QUALITY OF IMPACT EVALUATIONS AT USAID

Contracted under AID-OAA-M-13-00017

E3 Analytics and Evaluation Project

Prepared by:

Irene Velez, Team Leader, Management Systems International

DISCLAIMER
The authors' views expressed in this report do not necessarily reflect the views of the United States Agency for International Development or the United States Government.

(THIS PAGE INTENTIONALLY LEFT BLANK)

ABSTRACT

This study assessed the quality of the 133 USAID-funded impact evaluation (IE) reports published from fiscal years 2012 to 2019. The review started with an initial rating to determine if the IE report met USAID's IE definition. The rest of the review scored whether those IEs included quality elements across six domains that generate confidence in the reported results, including: sample size considerations, conceptual framing, treatment characteristics and outcome definitions and measurement, data collection and analysis, common threats to validity, and reporting of findings. The quality elements under each domain were separated into first- and second-tier scores to avoid arbitrary weights. The review found that 72 reports (54 percent) met USAID's IE definition, 37 reports (28 percent) did not provide a statistical justification for the validity of the comparison group, and 24 reports (18 percent) did not have a comparison group. Based on the first-tier score, 17 percent of reports were of high quality, 30 percent were of acceptable quality, and 8 percent were of lower quality. Based on the second-tier score, only 3 percent of reports were of high quality, 14 percent were of acceptable quality, and 28 percent were of lower quality. The review also found that 15 percent of IE reports conducted cost-effectiveness analysis. The study provided recommendations for improving IE quality, including reinforcing that IEs include a comparison group, updating guidance on IE reporting requirements, developing a standard IE report template and review checklist, conducting evaluability assessments, commissioning external peer reviews, integrating implementation fidelity monitoring into IE statements of work, including more information to disentangle and explain effects, and integrating ethical considerations and cost effectiveness as IE standards.

ACKNOWLEDGMENTS

The review team thanks Dr. Bhavani Pathak for her leadership as the USAID Contracting Officer's Representative for the E3 Analytics and Evaluation Project as well as Daniel Handel and Tania Alfonso from the Bureau for Policy, Planning, and Learning for their leadership and technical contributions. In addition, the team gives a special thanks to Joe Amick from the Global Development Lab for his feedback on the draft review instrument.

This report was a collaborative effort to which many individuals from Management Systems International contributed their time and expertise. Molly Hageboeck and Jacob Patterson-Stein contributed to the development of the review instrument and provided invaluable technical input for the conception of this study. Danielle Burke and Gary Glass provided feedback during the pretest and instrument finalization stage. Doug Krieger and Jeremy Gans conducted technical reviews of this report.

Irene Velez led this study, with Dan Killian and Idalia Rodriguez Morales serving as reviewers and team members. In addition, the study greatly benefitted from the support of Senior Project Manager Amanda Janczak and the professional editing expertise of Bettina Kimpton.

CONTENTS

TABLES

FIGURES

ACRONYMS

ADS	Automated Directives System
DEC	Development Experience Clearinghouse
E3	Bureau for Economic Growth, Education, and Environment (USAID)
FY	Fiscal Year
ICC	Intracluster Correlation Coefficient
IE	Impact Evaluation
LER	Office of Learning, Evaluation, and Research (USAID/PPL)
MDES	Minimum Detectable Effect Size
MSI	Management Systems International
PLC	Office of Planning, Learning, and Coordination
PPL	Bureau for Policy, Planning, and Learning (USAID)
QED	Quasi-Experimental Design
SOW	Statement of Work
USAID	United States Agency for International Development

EXECUTIVE SUMMARY

The Office of Planning, Learning, and Coordination in USAID's Bureau for Economic Growth, Education, and Environment (E3/PLC), in collaboration with the Office of Learning, Evaluation, and Research in USAID's Bureau for Policy, Planning, and Learning (PPL/LER), commissioned this examination of the quality of Agency-funded impact evaluations (IEs) published between fiscal years (FYs) 2012 and 2019. The review provides E3/PLC, PPL/LER, and other Agency evaluation advisors key findings on USAID IE strengths and weaknesses and provides recommendations to improve the quality of IE reports.

Study Methodology

USAID's 2011 evaluation policy and its 2016 update drew greater attention to IEs within the Agency. As the number of completed IEs has risen over the last few years, there has been a need for a formal review of their quality. This study assessed the 133 USAID-funded IE reports published between fiscal years 2012 to 2019 using a review instrument the review team designed and key staff from E3/PLC and PPL/LER reviewed. The review instrument was designed to be as objective as possible based on standards rather than subjective judgements of expert evaluators.

The review instrument starts with an **initial rating** that has two parts based on USAID's definition that IEs "require a credible and rigorously defined counterfactual." This initial rating assessed (1) whether the IE report had a comparison group, and (2) whether the IE report provided statistical justification for the validity of the comparison group before the intervention's start. If an IE report did not meet these two criteria, it did not proceed to receive a full review.

The rest of the review instrument is divided into six domains: sample size considerations, conceptual framing, treatment/intervention characteristics and outcome definitions and measurement, data collection and analysis, common threats to validity, and reporting of findings. Within each domain, the team included related questions and assessed whether the IE reports addressed that item fully, partially, or not at all. Some elements were simply yes or no. The elements within each domain were separated into a first tier and a second tier to avoid arbitrarily weighting the items differently based on perceived importance. Thus, the team assessed the IE reports based on two levels of quality:

- **First-tier score:** Consists of basic quality elements that are part of a standard IE report.
- **Second-tier score:** Consists of additional quality elements that are part of a credible IE (i.e., one generating confidence in the reported results) that can be used to make decisions.

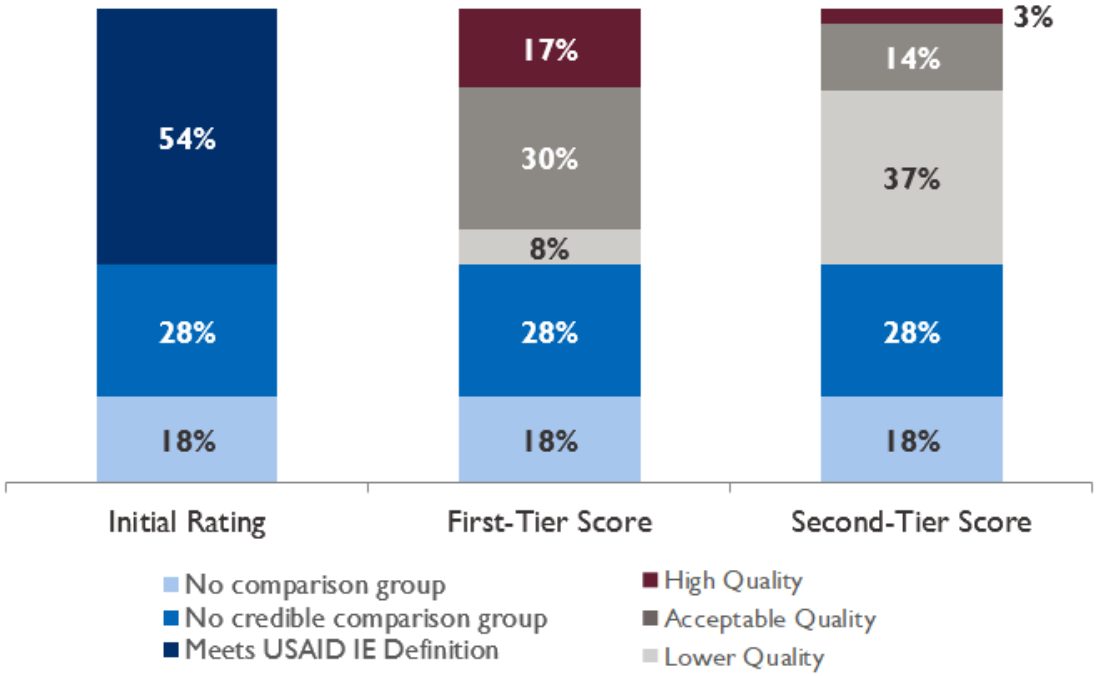
Key Findings

Initial Rating: Overall, the study found that 72 reports (54 percent) met USAID's IE definition (Figure 1). The remaining 61 reports (46 percent) did not meet USAID's IE definition because they did not have a comparison group (24 reports) or they did not provide a statistical justification for the validity of the comparison group (37 reports). These 61 reports did not proceed to a full review as their findings cannot be confidently attributed to the intervention that was evaluated. The study also found that while IE reports with a non-equivalent groups design made up 23 percent of the total reports, they accounted for only 3 percent of those that meet USAID's IE definition. IE reports with statistical matching or other statistical methods (e.g., instrumental variable, regression discontinuity design, synthetic control, geospatial panel framework) also made up 23 percent of the total reports but accounted for 31 percent of those that met USAID's IE definition. Over time, the percentage of IE reports that met USAID's IE definition increased substantially after the evaluation policy's release in 2011 (from 17 percent to a peak of 69 percent in FY 2018), but fell to 45 percent in FY 2019. Nonetheless, two to four IE reports were still published each FY that did not have a comparison group.

First-Tier Score: The study found that, based on the first-tier score, 17 percent of USAID-funded IE reports were of high quality, 30 percent were of acceptable quality, and 8 percent were of lower quality (Figure 1). There was a moderate association between first-tier score and FY (Cramer's V = 0.3807), as the percentage of high-quality IEs increased over time, peaked in FY 2018, and slightly decreased in FY 2019. The team found that most IE reports addressed quality criteria related to descriptions of the treatment and outcome measures as well as details about the data collection and analysis methods. The team also found that IE reports generally addressed criteria related to reporting findings. The team more often found limitations in criteria related to conceptual framing and reporting sample size calculations to ensure the IE's power to detect impact.

Second-Tier Score: The study also found that, based on the second-tier score, only 3 percent of USAID-funded IE reports were of high quality, 14 percent were of acceptable quality, and 28 percent were of lower quality (Figure 1). Unlike the first-tier score, the association between second-tier scores and FY was small (Cramer's V = 0.2184), due to the modest increase over time in the inclusion of additional quality elements that strengthen the credibility of IE findings so they can be used to make decisions. The team found that most IE reports did not address quality criteria at this level. Although most IE reports partially addressed quality criteria related to conceptual framing, the team found substantial limitations in addressing elements of sample size calculations, common threats to validity, and reporting of findings. These gaps present an opportunity to make the information included in IE reports more comprehensive to improve the findings' credibility so USAID stakeholders and external partners can use them for decision making.

FIGURE 1: OVERALL IE QUALITY REVIEW SCORES (N=133)



Cost Effectiveness: An additional area of interest for this study that was not factored into the quality scores was the use of cost-effectiveness analysis to link an intervention's effectiveness to its costs. The study found that 11 of the 72 IE reports reviewed (15 percent) conducted cost-effectiveness analysis of the activity/intervention using cost data and the impact estimates measured. While low, this figure is similar to the estimated percent of IEs with any value-for-money analysis at the World Bank (19 percent) and in the International Initiative for Impact Evaluation's (3ie's) IE repository (14 percent).

Summary Conclusions

Overall, the study findings show some quality improvements over time but critical gaps need to be addressed to improve the quality of USAID-funded IE reports. First, eight years after USAID's evaluation policy was put in place, there continues to be two to four IE reports published each year without a comparison group. These IE reports claim to be IEs but do not meet USAID's definition as they do not provide a comparison group to serve as a counterfactual to control for factors other than the activity/intervention that might account for observed change. Second, IE reports – in particular those with non-equivalent groups design – are not consistently providing statistical justification for the comparison group's validity. Providing this information ensures the reader that the evaluation findings can be attributed to the intervention evaluated, but more importantly it provides USAID stakeholders with an evaluability checkpoint. Third, quality elements that are not expected in performance evaluations were not usually included in the IE reports. However, not all quality elements that are explicit criterion in USAID's reporting guidance, and thus applicable to performance evaluations, were prevalent in IE reports. This implies a gap in guidance specific to IE reports and presents an opportunity for USAID to reinforce its general evaluation reporting criteria as well. Fourth, the study findings reveal the need for IE reports to shift from simply answering *whether* an activity/intervention is effective (e.g., only reporting impact estimates and statistical significance) to answering *why* there was an impact or lack thereof. The latter requires laying out a theory of change and defining outcome measures along the causal pathways, incorporating qualitative methods, conducting implementation fidelity monitoring, addressing common threats to validity, and providing explanations for null effects. Fifth, IE findings need to be more easily accessible to incorporate into practice. To accomplish this, IE reports need to include more discussion about the practical significance of impact estimates to interpret how large of an effect is the reported point estimate. This provides needed information for USAID stakeholders to decide whether the benefits are large enough to justify allocating resources in that activity/intervention. It also enables comparison to reported effects from other IEs. Finally, the study findings present an opportunity for USAID to include ethical considerations into evaluation guidance and reporting requirements and to operationalize the evaluation policy's call for cost-effectiveness.

Recommendations

Based on the study results, the team recommends the following actions for USAID to improve IE report quality to make IE findings more accessible and useful:[1]

1. PPL/LER and USAID evaluation managers should reinforce that, per USAID's evaluation policy, IEs must include a comparison group.
2. PPL/LER should provide updated, detailed guidance on the following specific elements that should be included in final IE reports:
    - Statistical justification of the comparison group's validity;
    - Explicit evaluation questions linked to the evaluation purpose;
        i. A theoretical framework including a literature review, a theory of change, and specific hypotheses to be tested;
    - Defined and operationalized outcomes in the methodology section, before presenting findings;
    - Specific power calculation parameters;
    - Detailed and complete information on common threats to validity;
    - Reporting findings that include point estimates and statistical significance as well as the control group mean to interpret the effect's magnitude;

[1] USAID released an update to Automated Directives System (ADS) 201 on October 28, 2020, after this report had been drafted. The ADS 201 revisions align with some of the recommendations outlined in this report.

- Discussion of null effects; and
- Actionable recommendations that advise USAID decisionmakers on the implications of the IE's results.

3. PPL/LER should develop a standard IE report template and review checklist to minimize the omission of important quality elements.
4. USAID evaluation managers should conduct evaluability assessments to ensure that only IEs that meet USAID's IE definition (i.e., adequately powered study to measure changes with a valid comparison group) are funded.
5. USAID evaluation managers should commission external peer reviews to assess the quality of evaluation designs and final reports, especially when there are gaps in internal technical capacity to adequately do so.
6. USAID evaluation managers should integrate implementation fidelity monitoring into IE statements of work.
7. PPL/LER should provide guidance to shift IEs toward reporting more information to disentangle and explain effects.
8. USAID should integrate ethical considerations as an IE standard to align with its Scientific Research Policy.
9. USAID should integrate the evaluation policy's call for cost-effectiveness as an IE standard.

INTRODUCTION

The United States Agency for International Development's (USAID's) Office of Planning, Learning, and Coordination in the Bureau for Economic Growth, Education, and Environment (E3/PLC), in collaboration with the Office of Learning, Evaluation, and Research in the Bureau for Policy, Planning, and Learning (PPL/LER), requested that the E3 Analytics and Evaluation Project[2] examine the quality of recent USAID-funded impact evaluations (IEs). This review assessed the quality of USAID IE reports published between fiscal years (FYs) 2012 and 2019. The review team assessed these reports using a review instrument based on standards rather than subjective judgements of expert evaluators so that it would be as objective as possible. The review provides E3/PLC, PPL/LER, and other Agency evaluation advisors key findings on USAID IE strengths and weaknesses and provides recommendations to improve the quality of IE reports. Annex A provides USAID's approved statement of work (SOW) for this review.

BACKGROUND

USAID's 2011 evaluation policy and its 2016 update[3] drew greater attention to IEs within the Agency. At the same time other organizations, including the Abdul Latif Jameel Poverty Action Lab (J-PAL) and the World Bank, were already demonstrating the feasibility of conducting experimental and quasi-experimental studies to examine development assistance outcomes across a range of sectors in developing countries.

Despite IEs' importance as a decision-making tool for foreign assistance programs, there has been little formal review of completed IE reports to ensure their consistency with Agency policy and guidance and with professional norms for IE quality. A 2013 meta-evaluation that MSI also conducted showed some quality improvements of USAID evaluations between 2009 and 2012, but only 3 percent of its sample (11 evaluations) were IEs; the remaining 97 percent were performance evaluations.[4] Similarly, a U.S. Government Accountability Office report that looked at evaluations across U.S. agencies completed in FY 2015 found that 26 percent of USAID evaluations were of high quality, 49 percent were of acceptable quality, and 26 percent were of lower quality. However, only 14 of the 63 USAID evaluations reviewed were IEs.[5]

DEFINING "IMPACT EVALUATION"

"Impact evaluations measure the change in a development outcome that is attributable to a defined intervention; impact evaluations are based on models of cause and effect and require a credible and rigorously defined counterfactual to control for factors other than the intervention that might account for observed change. Impact evaluations in which comparisons are made between beneficiaries that are randomly assigned to either a treatment or control group provide the strongest evidence of a relationship between the intervention under study and the outcome measured."

---

[2] Management Systems International (MSI), a Tetra Tech company, is the lead implementer of the E3 Analytics and Evaluation Project, along with team partners Palladium and NORC at the University of Chicago.

[3] USAID lays out its evaluation policies in its Automated Directives System (ADS). In September 2106, USAID issued a fully revised ADS 201 that addressed evaluation guidance, planning, and implementation and makes IEs a requirement for any activity within a project involving untested hypotheses or demonstrating new approaches that are anticipated to be expanded in scale or scope through U.S. government foreign assistance or other funding sources. Any activity or project designated as a "pilot" or "proof of concept" falls under this requirement. See https://www.usaid.gov/sites/default/files/documents/1870/201.pdf.

[4] Hageboeck, Molly, Micah Frumkin, and Stephanie Monschein. "Meta-Evaluation of Quality and Coverage of USAID Evaluations 2009–2012." Management Systems International, August 2013. https://pdf.usaid.gov/pdf_docs/PDACX771.pdf.

[5] U.S. Government Accountability Office. "Foreign Assistance: Agencies Can Improve the Quality and Dissemination of Program Evaluations." U.S. Government Accountability Office, March 2017. https://www.gao.gov/assets/690/683157.pdf.

Quality is critical to the utilization of evidence and learning generated by IEs. A 2016 study on the utilization of USAID evaluations that MSI also conducted found a significant relationship between evaluation utilization and average evaluation report quality scores at the operating unit level.[6] This finding could indicate that operating units with stronger "evaluation culture" prioritize and invest in both high-quality evaluations and their use. A 2014 systematic review indicated that clarity, relevance, and reliability of research findings (or lack thereof) serve as both a facilitator and a barrier to the use of evidence.[7] As the number of completed USAID IEs has risen over the last few years, there is a need for a formal review of their quality.

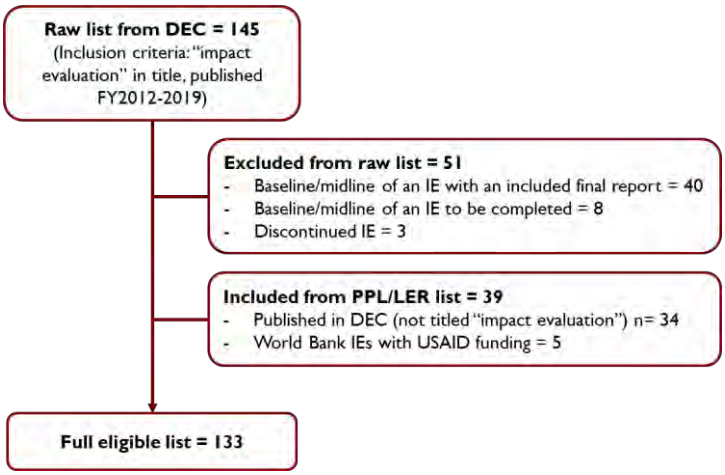
METHODOLOGY

SAMPLE

To assess the quality of USAID's IEs, the review team searched the Agency's Development Experience Clearinghouse (DEC) for all completed IE reports published between FYs 2012 and 2019. The team searched for all documents categorized as "final evaluation reports," "special evaluation reports," and "other USAID evaluations" that included "impact evaluation" in the title and were published in English. This search yielded 130 documents published between FYs 2012 and 2019.

The team conducted a second search on the DEC for documents categorized as "other USAID-supported study/document," "journal article," "evaluation summary," and "assessment" that included "impact evaluation" in the title (for the same time period). Most of these documents were design reports, baselines, midlines, or other types of assessments, but the search yielded 15 more documents, for an initial total of 145 documents. Of these 145 documents, the team excluded 51 documents that it identified as duplicates or baseline or midline reports for IEs that will be completed later.[8] This reduced the list to 94 reports.

FIGURE 2: SCREENING PROCESS FOR IE



PPL/LER then shared with the team a list of IEs it had identified. The team cross-checked this list with that of the evaluation reports it had generated from the DEC, yielding 39 additional IE reports. Thus, the number of eligible IEs included in this study is 133 reports published between FYs 2012 and 2019. Figure 2 summarizes this screening process and Annex B lists the 133 IE reports.

---

[6] Hageboeck, Molly, Micah Frumkin, Jenna L. Heavenrick, and Lala Kasimova. "Evaluation Utilization at USAID". MSI, February 2016. https://pdf.usaid.gov/pdf_docs/pa00kxvt.pdf.

[7] Oliver, Kathryn, Simon Innvar, Theo Lorenc, Jenny Woodman, and James Thomas. "A systematic review of barriers to and facilitators of the use of evidence by policymakers." *BMC Health Services Research* 14, no. 1 (2014): 2.

[8] Of the 51 excluded documents, 40 were baseline or midline reports of a completed evaluation that had a final report already included, 8 were baseline or midline reports of ongoing evaluations that will be completed later, and 3 were discontinued IEs that had not been completed.

REVIEW INSTRUMENT

The team developed the review instrument based on social science standards for research (the What Works Clearinghouse Procedures and Standards Handbook and the 2010 Consolidated Standards of Reporting Trials Statement). The team also drew from multiple USAID sources, including its evaluation policy, the 2013 Technical Note on Impact Evaluation, Automated Directives System (ADS) 201, and other evaluation reporting requirements. All the quality elements in the review instrument are suggested in USAID's 2013 Technical Note on Impact Evaluations but are not expected in Agency performance evaluations, so many are not part of USAID's evaluation reporting requirements. Once the team identified all the quality elements, it drafted the review instrument to assess whether the IE reports included these elements and removed subjective judgement about the adequacy of those elements. For example, the instrument asks whether the IE reports include several parameters for power calculations, but does not judge the adequacy of the power calculations. An IE report that includes these quality elements provides credible information on which to make decisions. This approach was also taken so that expert judgement is not needed to use the instrument, to encourage future application of this tool. Thus, the instrument can be shared with USAID evaluation managers and evaluation contractors as a tool for drafting and reviewing IE reports. The instrument also includes descriptive questions to gather information about the evaluations, such as country, technical sector, evaluation type, and additional details related to the quality elements.

The team pretested the review instrument by having the three reviewers assess a random subset of nine evaluations and then revised the instrument for clarity and adequacy based on the pretest results. E3/PLC and PPL/LER reviewed and approved the final review instrument (see Annex C).

The review instrument starts with an **initial rating** that has two parts based on USAID's evaluation policy and ADS 201, which defines IEs as evaluations that "measure the change in a development outcome that is attributable to a defined intervention. Impact evaluations are based on models of cause and effect and require a credible and rigorously defined counterfactual to control for factors other than the intervention that might account for the observed change." This initial rating assessed whether the IE report provided statistical justification for the validity of the comparison group before the intervention's start. If an IE report did not meet this criterion, it did not proceed to receive a full review.

For the full review, the review instrument is divided into six domains: sample size considerations, conceptual framing, treatment characteristics and outcome definitions and measurement, data collection and analysis, common threats to validity, and reporting of findings (Table 1). Within each domain, the team included related questions to assess the quality elements, which were separated into a **first tier** and a **second tier** to avoid arbitrarily weighting the items differently based on perceived importance. Thus, the team assessed the IEs based on two levels of quality:

- **First-tier score:** Consists of basic quality elements that are part of a standard IE report.
- **Second-tier score:** Consists of additional quality elements that are part of a credible IE (i.e., one generating confidence in the reported results) that can be used to make decisions.

TABLE 1: IE QUALITY SCORING ELEMENTS

Domain	First-Tier Score: Basic Quality Elements	Second-Tier Score: Quality Elements for Credible IEs
Sample Size and Power Calculations	Sample size	Expected take-up in power calculations
	Power level	Expected attrition in power calculations
	Minimum detectable effect size	Cluster level defined
		Intracluster correlation coefficient
Conceptual Framing	Evaluation purpose and intended use	Literature review
	Impact evaluation questions	Local context
	Theory of change	Evaluation hypotheses
Treatment and Outcome Measures	Treatment description	
	Outcome measures description	
Data Collection and Analysis Methods	Data collection methods description	
	Data analysis methods description	
Common Threats to Validity		Treatment fidelity
		Actual treatment take-up
		Contamination across groups
		Actual attrition
		Non-response/missing data
Reporting Findings	Statistical significance of impact estimates	Practical significance of effect sizes
	Connect findings to recommendations	Explanation of null effect

REVIEW PROCESS

The review team took several steps to ensure consistency among reviewers' responses. First, the team leader trained two additional reviewers on the review instrument and provided a rater's guide. Then, the team leader conducted a calibration exercise during the pretest with the other two reviewers to ensure high inter-rater reliability. The three reviewers reviewed the same nine IE reports on their own. The team then met to discuss how each element was coded and compared answers. Where there was disagreement, the team discussed and agreed on how it should be coded.

Following the pretest, the three reviewers applied the review instrument to each IE report in the final sample. The team leader conducted a second independent review of a random subset of reviewed reports to conduct inter-reliability checks throughout the review process. The team leader provided feedback to the reviewers on any answers where there was disagreement and subsequently harmonized responses. The team leader also held weekly calls with reviewers and provided remote support as needed. In addition, the team leader conducted a second review of all reports that did not meet the initial rating criteria.

SCORING AND ANALYSIS

The team assessed whether the IE report addressed each quality element. For each quality element, if the IE report addressed the item fully it received one point, if it addressed the item partially it received half a point, and if it did not address the item at all it received zero points. For the elements with yes and no responses, if the IE report addressed the item it received one point and if it did not address the item at all it received zero points. The team then summed the scores for the individual quality elements under each tier to generate the absolute first-tier and second-tier scores. Since the elements specific to cluster designs are not relevant to all IEs, the team divided the absolute scores by the total number of relevant quality elements, resulting in final percentage scores (0 to 100 percent).

For the first-tier score, the team rated IE reports as high quality if they received a final score of 75 percent or more, acceptable quality if they received a final score between 50 and 74 percent, and lower quality if they received a final score less than 50 percent. For the second-tier score, the team rated IE reports as high quality if they received a final score of 66 percent or more, acceptable quality if they received a final score between 50 and 65 percent, and lower quality if they received a final score less than 50 percent. The study only reports the categorical score rather than the percentage score because the difference between 55 and 60 percent, for example, is not particularly meaningful. The team ensured that the cutoff threshold did not negatively affect the categorical score distribution, so that there was not a high number of IEs scoring right under the threshold. In addition, the distribution of the first-tier and second-tier scores overlap in an intuitive way. Since the first-tier score represents the achievement of basic quality elements, the second-tier score does not result in a higher category than the first-tier score (Table 2).

TABLE 2: DISTRIBUTION OF QUALITY SCORES

First-Tier Scores	Second-Tier Scores		
	Lower Quality	Acceptable Quality	High Quality
Lower Quality	100%		
Acceptable Quality	68%	32%	
High Quality	55%	27%	18%

Note: Percentages correspond to row frequencies.

The team analyzed the responses to the review instrument using Stata16 to produce descriptive statistics of the overall quality scores and quality elements (i.e., frequency distributions) and to measure the strength of association between quality and evaluation type and FY (Cramer's V). Since the study includes all the identified IE reports (instead of selecting only a sample), the team did not include chi-square to determine significance. Following discussions with E3/PLC and PPL/LER, results for individual reports will not be made public.

STRENGTHS AND LIMITATIONS

This study has several strengths and limitations. Since the study includes all USAID-funded IEs the team could identify since the establishment of the Agency's evaluation policy, the findings are reflective of the status of IEs at USAID and do not face issues of generalizability that have limited other assessments of evaluation quality. The study's approach to assessing evaluation quality as the inclusion of specific quality elements in the IE report led to robust inter-rater reliability. In addition, jointly developing and testing the instrument led to substantial agreement between reviewers even though reviews were conducted independently. These strengths increase the credibility of the study results.

The study also has a few limitations, including:

- The approach to assess the inclusion of quality elements in the IE report does not allow the study to make conclusions on the validity of the IE design or reliability of the IE findings. Thus, the study focuses on the inclusion of quality elements as a proxy for IE quality.
- The reviewers applied the instrument only to information provided in the final IE reports. The study did not examine any other evaluation documentation, such as design reports or baseline reports. Thus, some IEs may have been assessed negatively for not including information in the final IE report that may have been included in a previous evaluation document. The exception to this was when the final IE report mentioned that baseline balance tests had been previously

conducted. In these cases, the reviewers searched for the baseline report and rated this item based on the information found in the baseline report if it was publicly available.

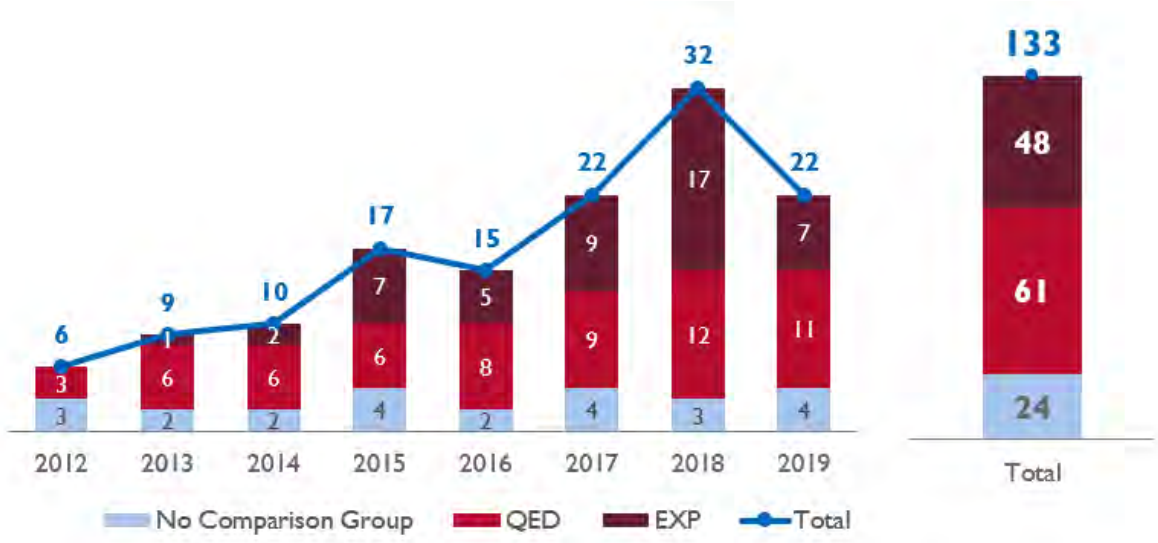
- Some of the information assessed might not have been included in the IE report since it was not a part of the original evaluation SOW (e.g., the evaluation questions). Thus, some evaluations may have been assessed negatively for not including information that was not provided to the evaluation team by the commissioning USAID operating unit.
- The quality scores do not include important elements that are not required or recommended by USAID guidance such as ethical considerations and cost effectiveness. Nonetheless, the study assessed these elements and reports on them, but does not include them in the quality scores.

FINDINGS

OVERVIEW OF IE REPORTS AT USAID

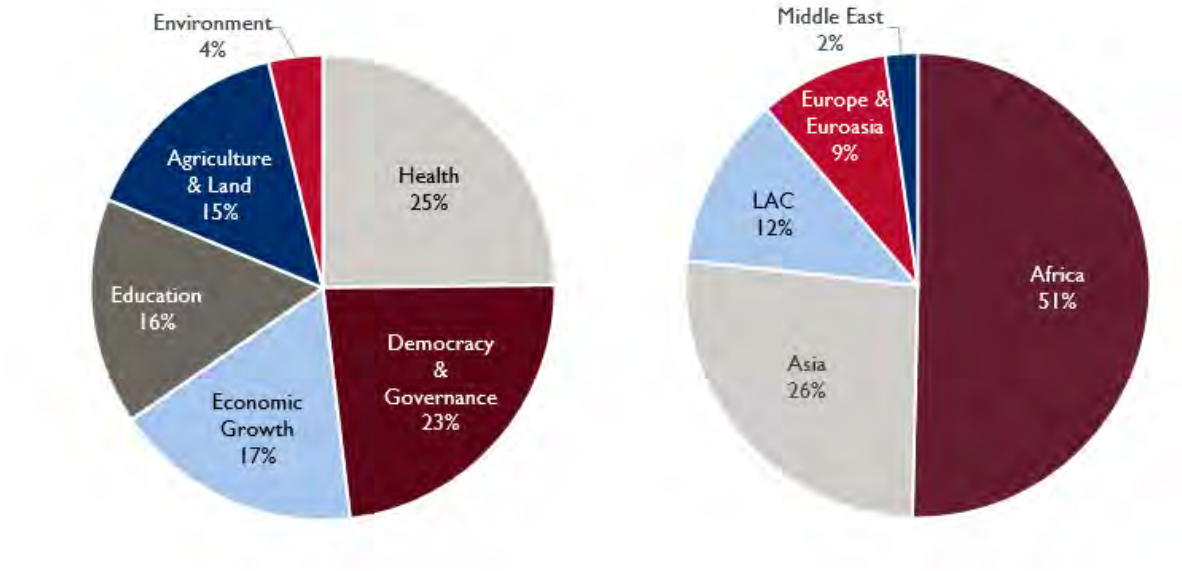
The number of USAID-funded IE reports published each FY has risen since 2012, following the evaluation policy's issuance in 2011, and peaked in 2018 (Figure 3). The decline in the number of IE reports for FY 2019 was due to a drop in IE reports with an experimental design, which had grown substantially in the previous years. The number of IE reports with quasi-experimental designs grew steadily during this entire time period and IE reports that did not have a comparison group remained consistently at two to four per year.

FIGURE 3: NUMBER OF IE REPORTS, FY 2012 – 2019 (N=133)



Of the 133 IE reports included in this review, 51 percent took place in Africa, 26 percent in Asia, 12 percent in Latin America and the Caribbean, 9 percent in Europe/Eurasia, and 2 percent in the Middle East (Figure 4). One-quarter of the IE reports were in the health sector and almost another quarter (23 percent) were in the democracy and governance sector; the latter being a sector in which IEs are newer than in more traditional sectors such as education and economic growth. Approximately the same amount of IE reports were from the sectors of economic growth (17 percent), education (16 percent), and agriculture and land (15 percent). Four percent of IE reports were from the environment sector.

FIGURE 4: SECTOR AND GEOGRAPHIC DISTRIBUTION OF THE IES (N=133)



Forty-eight IE reports described using experimental designs consisting of randomized assignment to treatment employed mostly through stratified randomization (38 percent), simple randomization (29 percent), matched pairs (17 percent), and blocked randomization (13 percent). Of the experimental IEs, only two reports did not provide sufficient information to determine how randomization was conducted. In addition, all 48 experimental IE reports randomized assignment to treatment at the same unit level as that of the intervention's implementation. Sixty-one IE reports described quasi-experimental designs in which the evaluator manipulates group assignment somehow but not through a randomized process. The most common quasi-experimental design was non-equivalent group design (49 percent), where the comparison group was either hand matched on a limited number of observable characteristics (10 percent) or selected through convenience assignment because it was nearby or feasible to reach (39 percent). One third of quasi-experimental IE reports conducted statistical matching such as propensity score matching and 18 percent used other statistical methods (e.g., instrumental variable, regression discontinuity design, synthetic control, geospatial panel framework). The remaining 24 IE reports did not have a comparison group and are described in more detail in the next section.

INITIAL RATING

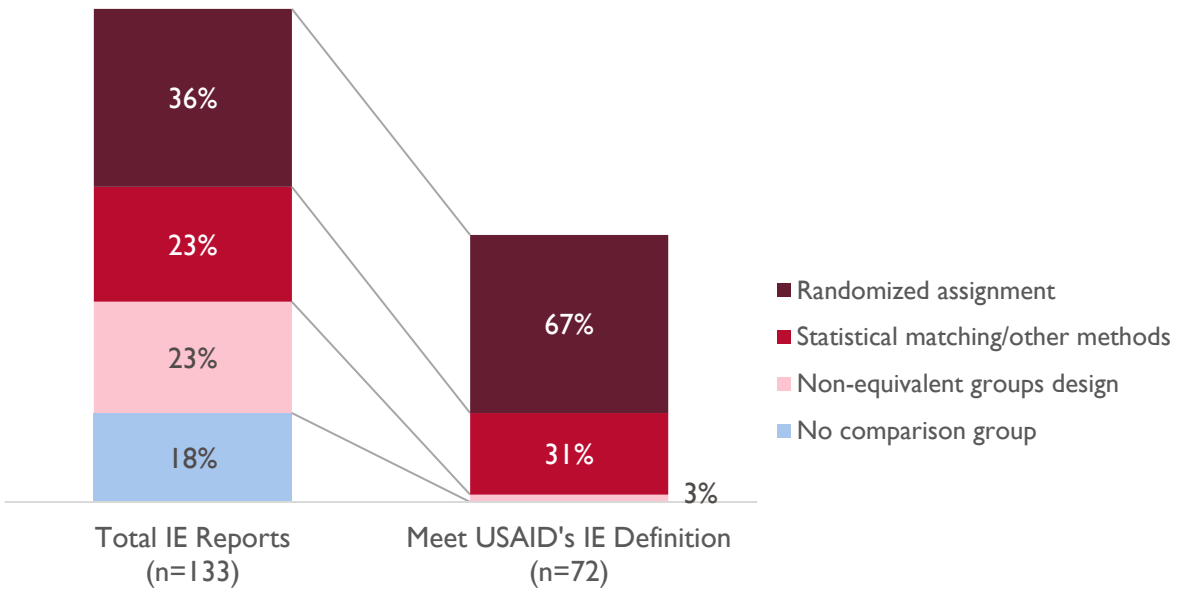
Each of the 133 IE reports went through an initial rating to determine whether, based on the IE report, it meets USAID's evaluation policy definition of an IE. This determination requires assessing (1) whether the IE report described a control/comparison group, and (2) whether the IE report provided a statistical justification that the control/comparison group is a good comparison for the treatment group. The review found that 24 IE reports did not have a comparison group and that 37 IE reports did not provide any statistical justification for the validity of the comparison group.[9] The remaining 72 IE reports (54 percent) provided information to show that the treatment and control groups were highly similar in key characteristics prior to the intervention (Figure 6). These 72 IE reports consisted of all 48 experimental

[9] In some cases, the final IE report mentioned that baseline balance tests had been previously conducted. The reviewers searched for the baseline report and rated this item based on the information found in the baseline report. However, for six IEs the baseline report was not found, so these were rated as not providing statistical justification for the comparability of the comparison group.

IEs, which satisfied the initial rating criteria by default;[10] 13 quasi-experimental IE reports that included a table with balance tests across various variables to statistically show baseline equivalence between the treatment and control group; and 11 quasi-experimental IE reports that did not have a table but included other analysis to justify comparability.

Since quasi-experimental IEs can use various design methods, the team analyzed the initial rating by these design methods. The team found that while IE reports with a non-equivalent groups design make up 23 percent of the total reports, they accounted for only 3 percent of those that met USAID's IE definition. IE reports with statistical matching or other statistical methods (e.g., instrumental variable, regression discontinuity design, synthetic control, geospatial panel framework) also made up 23 percent of the total IE reports but accounted for 31 percent of those that met USAID's IE definition (Figure 5).

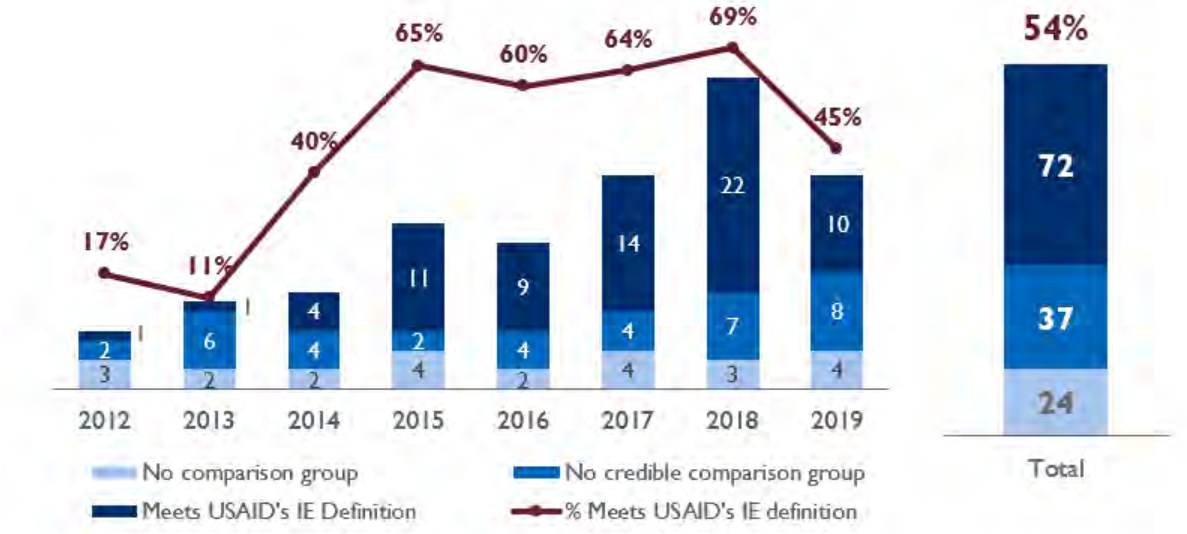
FIGURE 5: INITIAL RATING BY DESIGN METHOD



The percentage of IE reports that met USAID's IE definition rose substantially after the evaluation policy's establishment in 2011 and remained consistent between FY 2015 and FY 2018 until falling in FY 2019 (Figure 6). However, there continues to be two to four IE reports published each year without a comparison group and the number of IE reports that do not provide statistical justification for the comparison group started to increase again in the past two years.

[10] Experimental IEs were exempt from this screening because "randomized assignment of treatment ensures that systematic differences between groups do not drive differences in outcomes. Randomized assignment to treatment is the most effective tool for eliminating selection bias because it removes the possibility of any individual characteristic influencing selection. Because units are not assigned to treatment or control groups based on specific characteristics but rather are divided randomly, all characteristics that might lead to selection bias, such as motivation, poverty level, or proximity, will be roughly equally divided between the treatment and control groups." (USAID, Impact Evaluations Technical Note) The screening used for experimental evaluations was that they randomized assignment to treatment at the same unit level as that of the intervention's implementation. All experimental evaluations met this criterion.

FIGURE 6: INITIAL RATING BY FISCAL YEAR (N=133)



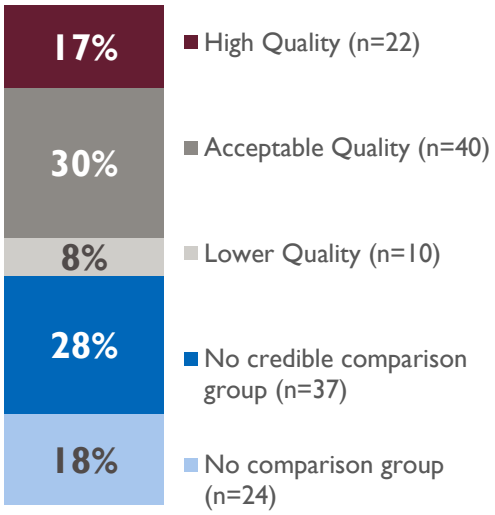
Note: The bar graph shows the total IE reports published each FY by initial rating. The line reflects the percentage of IE reports that meet USAID's IE definition each FY. The right bar graph and percentage refer to the total IE reports published between FYs 2012 and 2019.

As discussed in the methodology section, the rest of the review consists of a two-tiered score. The first tier consists of basic elements that need to be included in a standard IE report. The second tier consists of additional quality elements that are part of a credible IE (i.e., one generating confidence in the reported results) that can be used to make decisions. Annex D shows the findings for each quality element across the two tiers.

FIRST TIER: BASIC QUALITY ELEMENTS

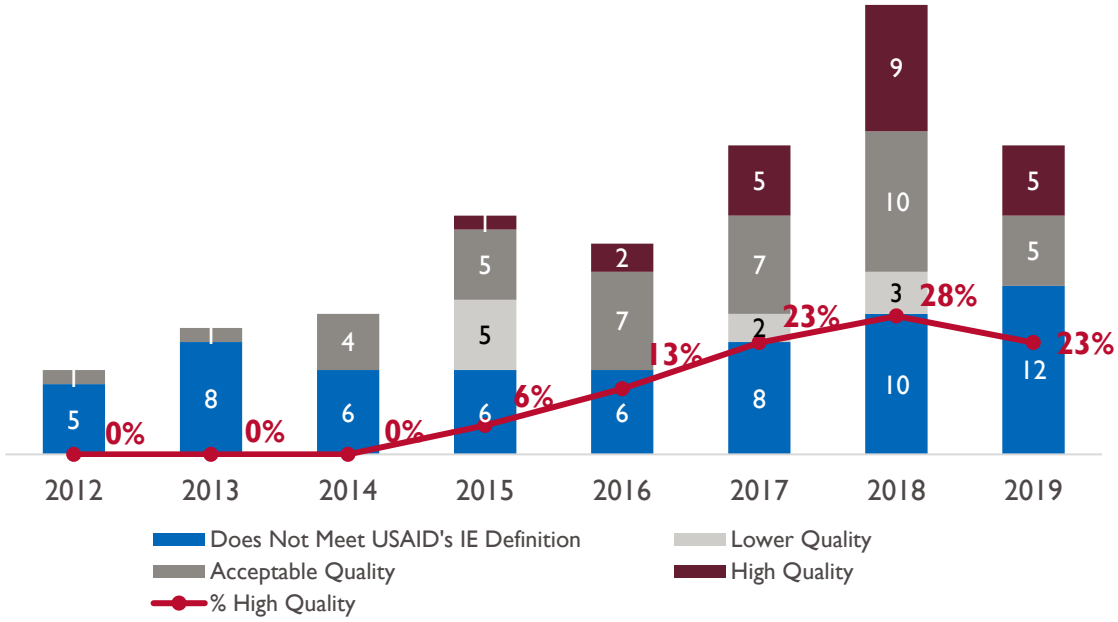
Overall, 17 percent of the IE reports rated as high quality (Figure 7), meaning they addressed or partially addressed at least 75 percent of the basic quality elements in the first tier (Table 1). Thirty percent of the IE reports were of acceptable quality, meaning they addressed or partially addressed between 50 to 74 percent of the elements in the first tier. Eight percent of the IE reports were of lower quality, having addressed or partially addressed less than half of the elements in the first tier. The remaining 46 percent of IE reports did not meet USAID's IE definition. The 61 IE reports that did not meet the initial rating criteria were excluded from the full review because their findings could not be confidently attributed to the intervention evaluated. The remaining first-tier findings in this section correspond to the 72 IEs that meet USAID's IE definition.

FIGURE 7: FIRST-TIER SCORE (N=133)



The team found a moderate association between first-tier score and FY (Cramer's V = 0.3807), as the percentage of high-quality IEs increased over time, peaked in FY 2018, and slightly decreased in FY 2019 (Figure 8).

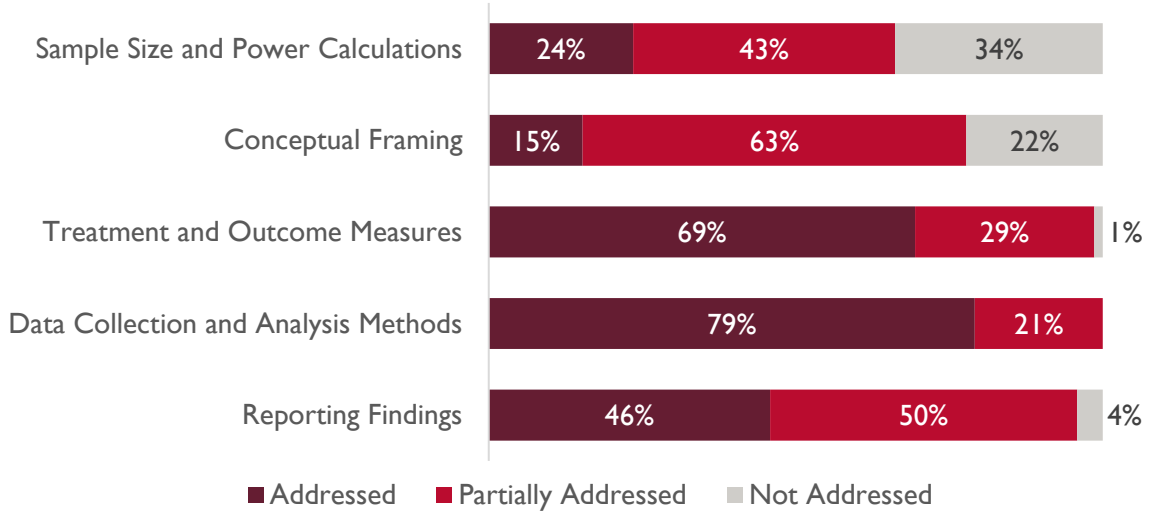
FIGURE 8: FIRST-TIER SCORES BY FISCAL YEAR (N=133)



Note: The bar graph shows the IEs by first-tier score and the line reflects the percentage of IEs with high quality first-tier scores each FY.

Figure 9 breaks down the first-tier score by domain. The team found that most IE reports addressed quality criteria related to descriptions of the treatment and outcome measures as well as details about the data collection and data analysis methods. The team also found that IE reports generally addressed criteria related to reporting findings. The team more often found limitations in criteria related to conceptual framing and reporting sample size calculations to ensure the power of a study to detect impact.

FIGURE 9: SUMMARY OF BASIC QUALITY ELEMENTS (N=72)



Sample Size and Power Calculations: Although these 3 basic quality elements ensure that the evaluation is set up to detect changes in outcome measures and that null effects are not a result of an underpowered study, only 24 percent of IE reports addressed all 3 elements, 43 percent addressed some of these elements (9 percent addressed 2 and 34 percent addressed only 1), and 34 percent did not address any of these elements.

Conceptual Framing: Only 15 percent of IE reports addressed all 3 basic quality elements to set up the evaluation's conceptual framing. Sixty-three percent addressed some of these elements (28 percent addressed 2 and 35 percent addressed only 1), and 22 percent did not address any of these elements.

Treatment and Outcome Measures: These two basic quality elements provide information about what is being evaluated and how it is measured. Sixty-nine percent of IE reports provided a clear description of both treatment and outcomes measures, 29 percent provided a clear description of only 1 of these 2 elements, and 1 IE report did not address either.

Data Collection and Analysis Methods: These two basic quality elements provide information about how the IE collected and analyzed data. Seventy-nine percent of IE reports provided a clear description of both data collection and analysis methods, and 21 percent provided a clear description of only 1 of these 2 elements.

Reporting Findings: These two basic quality elements make evaluation findings and their learning more accessible so it is easier for USAID officers to use the evidence to inform decision making. Forty-six percent of IE reports addressed both elements, 50 percent addressed 1 of these elements, and 4 percent did not address either.

SAMPLE SIZE AND POWER CALCULATIONS

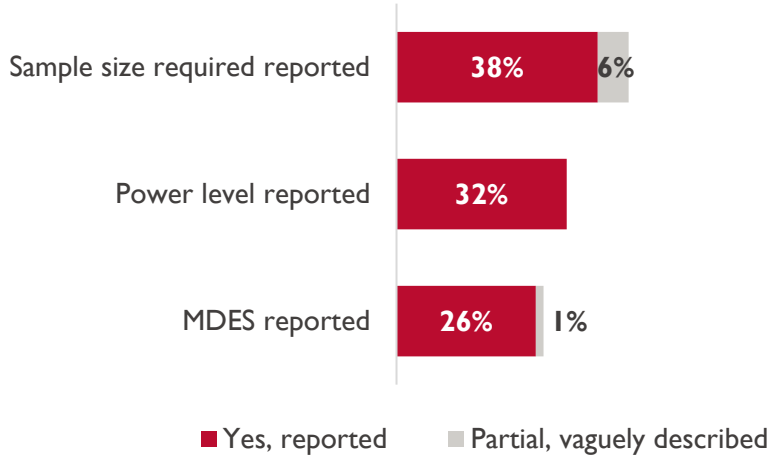
A critical element of IE quality is the use of power calculations to determine the adequate sample size to confidently detect meaningful treatment effects. This serves both quality and cost-efficiency purposes. First, it ensures that the evaluation will be able to detect a program impact if there truly is one or, put another way, it minimizes the risk of concluding that a program has had no impact when it has in fact

had one (false negative). Second, it ensures that only the necessary data are collected and that resources are used efficiently (data collection usually constitutes the largest cost of an IE).

Underpowered IEs have serious practical consequences. For example, if the IE were to conclude incorrectly that the program was not effective, implementers or donors may choose to reallocate resources elsewhere, thereby preventing continuation of positive impact. Conducting and reporting on power calculations allows the reader to confirm that null effects are not due to the study being underpowered. However, only one-third of IEs reported their power calculations. An additional one-fifth mentioned conducting power calculations in the design stage but did not include the information in the final report. Almost half did not mention or include power calculations in the report.[11]

In addition, only 38 percent of IEs reported the sample size needed to ensure adequate power, 32 percent reported the power level used, and 26 percent reported the minimum detectable effect size (MDES) (Figure 10). Of the 18 IEs that reported the MDES, only 8 (44 percent, or 12 percent of the IEs reviewed) explained the basis for the MDES used, such as project targets, results from similar studies or programs, ex-ante simulations, or the cost-benefit assessment threshold (the smallest effect that would make it worthwhile to run the program rather than dedicating resources elsewhere). As MDES selection is not a technical decision, reporting its basis can help readers understand what was meaningful to the parties involved.

FIGURE 10: SAMPLE SIZE AND POWER CALCULATIONS (N=68)



Note: This graph excludes the four IEs that relied on extensive administrative data (with thousands of observations) as their primary data source, obviating the need for ex-ante power calculations.

CONCEPTUAL FRAMING

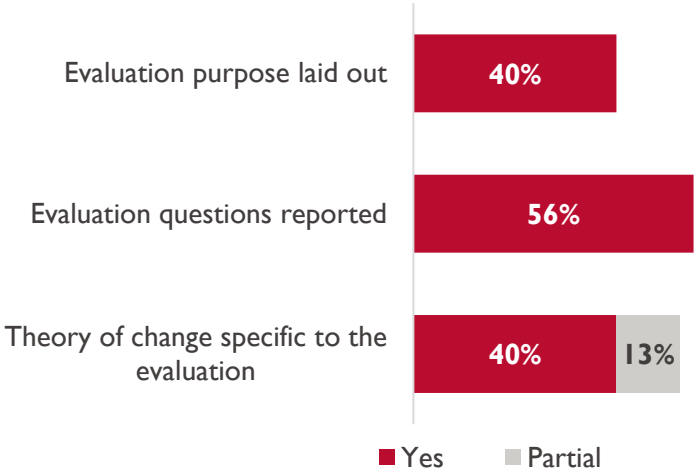
In general, for an evaluation to be useful and relevant it must explicitly link evaluation questions to the evaluation's purpose and intended uses so that USAID staff, partner governments, and other key stakeholders can make informed decisions. The reader can expect to learn more from evaluations that are designed with a clear purpose and development hypothesis. The development hypothesis should clearly define the logic of the intervention(s) being evaluated, with emphasis on the treatment (independent variable) and the principal anticipated outcomes (dependent variables). This theory of

[11] The conclusion is not that these IEs did not conduct power calculations ex-ante, only that they did not mention or include them in the final IE report. It is possible that power calculations were previously conducted; however, the reviewers did not search for design reports to confirm.

change, or logic model, also provides the basis for the questions to be addressed by the IE. Identifying key evaluation questions also improves the evaluation design's quality and guides evaluators to identify the right kinds of data and appropriate methodology for data collection and analysis. This study looked at these three elements of conceptual framing as part of the first-tier score: (1) an explicit evaluation purpose, (2) reported evaluation questions, and (3) a theory of change that is specific to the evaluation.

While all reviewed IE reports stated the evaluation objective was to test the effectiveness of an intervention, only 40 percent laid out the evaluation purpose describing the intended use(s) of the evaluation findings (Figure 11). These IE reports stated that the evaluation findings would inform future programming, provide evidence about the cost effectiveness of the intervention for scale-up or replication, or improve implementation mechanisms. This study also found that IE reports with an evaluation purpose section were almost three times more likely to describe not only the evaluation objective but also to include information about the evaluation's intended use(s).

FIGURE 11: CONCEPTUAL FRAMING ELEMENTS (N=72)



Although reporting the evaluation questions is an explicit criterion in USAID's evaluation policy to ensure the quality of an evaluation report, only 56 percent of IEs explicitly listed the evaluation questions in the report (Figure 11). These IEs had a median of 4 evaluation questions, with 80 percent of the IEs having 6 evaluation questions or fewer. However, 12 percent of the IEs listed between 10 and 20 evaluation questions. In addition, of the IEs that reported the evaluation questions, only 27 percent matched the evaluation questions in the SOW or, if they do not match, explained the deviation. Seven percent of IEs had evaluation questions that did not match the SOW and provided no explanation, but more importantly, 66 percent of IEs that reported the evaluation questions did not include the SOW as an annex to the report.

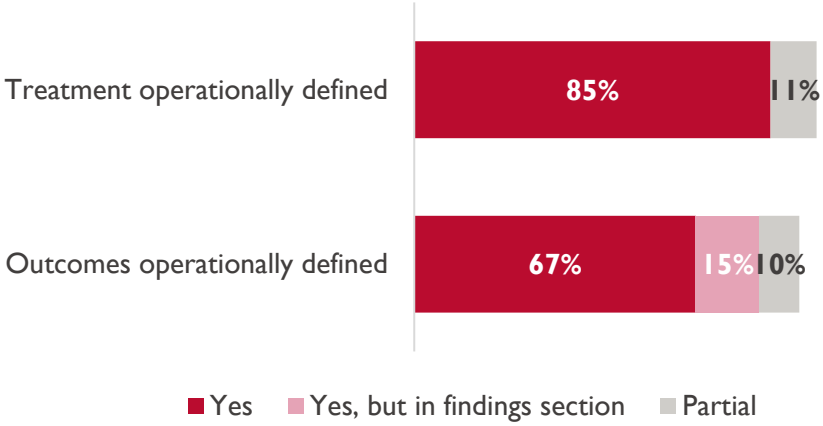
Only 40 percent of IEs had a graphic and/or narrative description of the theory of change, specific to the evaluation, that described the causal pathways through which the intervention to be examined was supposed to achieve the intended results. Another 13 percent presented a vague theory of change or a results framework focused on results beyond the scope of the evaluation (Figure 11).

TREATMENT AND OUTCOME MEASURES

Linked to the evaluation questions and theory of change, this study assessed whether the IE reports clearly described the activity/intervention to be evaluated and clearly defined the outcome measures. The treatment description plays a critical role in enabling the reader to understand what is being

evaluated and adequately use the evaluation's findings to design or implement new programs. The description should be detailed enough to support an effort to replicate the treatment; it therefore should include information such as what is being delivered, by whom (e.g., nurse, teacher, community agent), dosage, and frequency. Eighty-five percent of the IEs reviewed provided a detailed description of the treatment, and an additional 11 percent provided a vague or general description of the intervention. Only four percent of IEs did not provide a description of the treatment (Figure 12).

FIGURE 12: DEFINITIONS OF TREATMENT AND OUTCOME MEASURES (N=72)



Two treatment characteristics usually outside the control of the evaluator are complexity and uniformity. Although these two items are not included in the quality score, they do provide more information about what is evaluated. Fifty-one percent of IEs consisted of complex treatments with bundled interventions, which makes it difficult to distinguish which component is driving the effect, if any, and which components of the intervention would be needed to replicate the impact elsewhere (Figure 13).

FIGURE 13: TREATMENT COMPLEXITY (N=72)

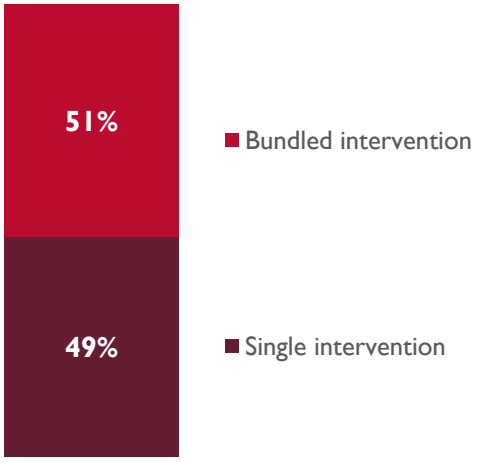
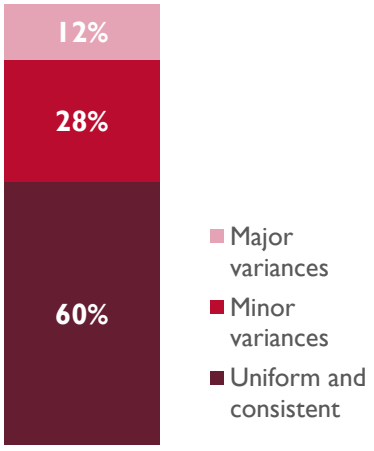


FIGURE 14: TREATMENT UNIFORMITY (N=72)



In addition, 40 percent of IEs consisted of treatments that had minor or major variances across units within the treatment group that were not part of the evaluation design or analysis (Figure 14). This lack of consistency and uniformity was due to having multiple implementing partners delivering the treatment

differently, restrictions imposed by local governments on what or how programs can be delivered, or logistical considerations or challenges that required modifications.

The description of outcome measures also plays a critical role in IE quality, particularly their specification, definitions, and metrics. For example, rather than stating that the outcome of interest is employment, the report should specify that the outcome of employment will be measured by whether an individual currently works for 20 hours or more per week (or another equivalent definition). This specification is even more important for broader constructs such as civic engagement, social cohesion, and empowerment, which can be defined and measured in various ways. Two-thirds of the IE reports described the outcome measures and their indicators in the methodology section before presenting the findings. However, 15 percent provided descriptions of the outcome measures in the findings section. An additional 10 percent provided a vague or general description of the outcome measures and 8 percent did not provide any description of the outcome measures (Figure 12). In addition, 90 percent measured outcomes once at baseline, prior to the start of the treatment, and 4 percent reported more than 1 round of pre-treatment measurements. Only six percent of IE reports reported no baseline. Fifty-seven percent of IEs reported 1 endline measurement, 24 percent had 2 rounds of post-treatment measurement, and 10 percent had 3 rounds of post-treatment measurement.

DATA COLLECTION AND ANALYSIS METHODS

Detailed descriptions of data collection and analysis methods enable readers to assess the evaluation findings' validity and reliability. Most IE reports addressed these basic quality criteria: 94 percent provided a clear and detailed description of the data collection methods (Figure 15), including specific information on how and from whom data were collected and a description of the instruments used to collect the data. Almost all IEs used survey data and the four IEs that did not use surveys relied on administrative data. Nonetheless, almost half of the IEs (49 percent) also used qualitative methods to supplement or complement quantitative data, with 46 percent conducting key informant interviews and 32 percent conducting focus group discussions. Forty-two percent of IEs used administrative data and 17 percent used other data collection methods, such as anthropometry data, direct observation, and student reading assessments. However, only 47 percent of IE reports included the data collection instruments in an annex.

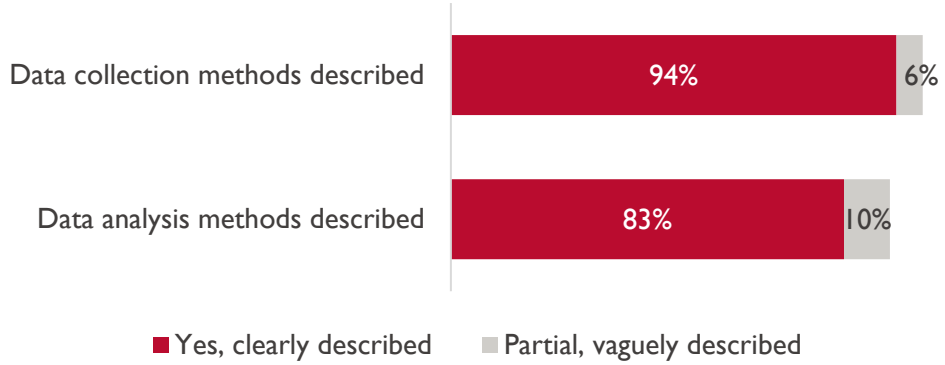
Eighty-three percent of IE reports provided a clear and detailed description of the data analysis methods, including model specifications and tests to determine impact estimates and/or group differences. Ten percent of IEs provided only a vague description, and 7 percent did not describe the data analysis methods at all (Figure 15). In addition, 64 percent conducted subgroup analysis to assess how the impact estimates differed by groups. Of these, 65 percent assessed how impact estimates differed by gender. Although disaggregation by gender is not a necessary quality element, USAID's evaluation policy states that if person-level outcomes or impact are assessed, they should also be separately assessed for males and females.

USAID's evaluation policy also calls for evaluations to be undertaken in a manner that ensures credibility, lack of bias, and transparency. One potential way to meet these requirements is through a pre-analysis plan. Although not required, a pre-analysis plan—written in advance of data collection—describing how the evaluator will analyze the data can help avoid data mining and specification searching by setting out in advance exactly the specifications that will be run and with which variables.[12] This review found that only 21 percent of IE reports included a pre-analysis plan (or design report with a detailed data analysis section) in the annex or referenced and provided a link to access it. Four percent

[12] B. A. Olken, "Promises and Perils of Pre-Analysis Plans." *Journal of Economic Perspectives* 29, no. 3 (2015), 61–80.

of IEs mentioned a pre-analysis plan vaguely but did not provide a means to access it, and the remaining 75 percent of IE reports did not mention a pre-analysis plan.

FIGURE 15: DATA COLLECTION AND ANALYSIS METHODS (N=72)



This study also looked at whether the IEs reported basic ethics considerations, such as research approval from an authorized institution (e.g., registered institutional review board or a government institution in the host country) and informed consent from respondents before the start of data collection. These two elements are not part of the quality scores but are important information to include in the IE report to ensure ethical conduct. Twenty-six percent of IEs reported receiving both ethics approval and informed consent, 31 percent reported receiving informed consent only, and 41 percent did not report on either ethics consideration.[13]

REPORTING FINDINGS

Clear reporting of findings also enables the reader to understand their significance and assess their validity. Elements such as number of observations, numerical reporting of treatment effects, and statistical significance are important. Although 88 percent of IEs reported sample sizes, not all of these reported the number of observations in the tables and narrative describing the treatment effects. Many IEs simply reported the sample size that resulted from data collection efforts. Most IEs (73 percent) reported effect sizes as absolute or relative changes in the outcome measure, but 26 percent of IEs included standardized effect sizes, which make it easier to understand the magnitude and practical significance of the treatment effect. Almost all IEs (92 percent) reported the statistical significance of treatment effects using p-values or confidence intervals. However, five percent of IEs reported the point estimate and stated that it was statistically significant without reporting the actual confidence interval or p-value, and three percent of IEs did not report on statistical significance at all (Figure 16).

Another element of reporting the findings is to connect them to conclusions or recommendations advising USAID decision makers on the implications of the IE's results for continuing/discontinuing, replicating, or scaling up the intervention(s) evaluated. USAID's commitment to evidence-based decision-making benefits from having evaluation reports that translate findings into actionable recommendations. However, only 50 percent of IE reports provided actionable recommendations on the implications of the IE results, 8 percent included recommendations but did not advise USAID decision makers on whether to continue or reallocate resources, and 42 percent did not include any recommendations (Figure 17).

[13] This excludes the four IEs that used only administrative data.

FIGURE 16: REPORTING STATISTICAL SIGNIFICANCE OF TREATMENT EFFECTS (N=72)

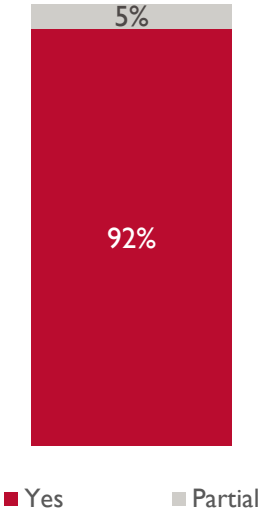
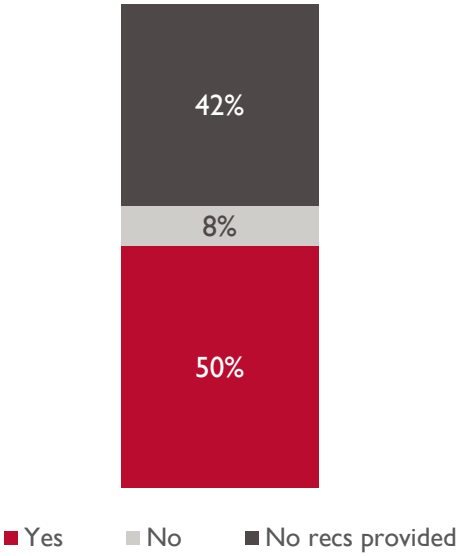


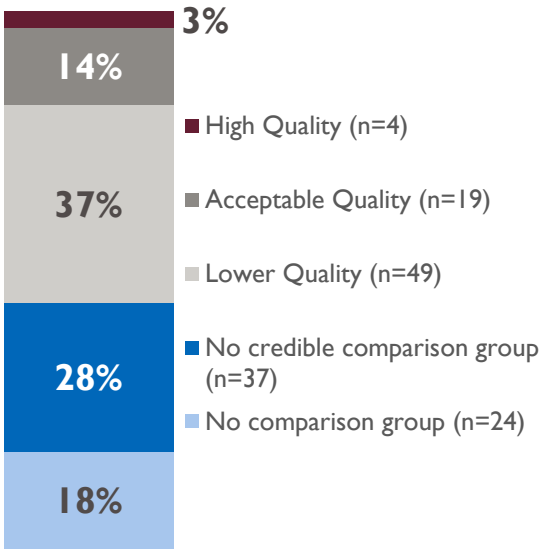
FIGURE 17: CONNECTING FINDINGS TO ACTIONABLE RECOMMENDATIONS (N=72)



SECOND TIER: QUALITY ELEMENTS FOR CREDIBLE IMPACT EVALUATIONS

This section discusses the second-tier score, which consists of additional elements that are part of a credible IE (i.e., one generating confidence in the reported results) that can be used to make decisions. Overall, only three percent of USAID IE reports were rated high quality (Figure 18), meaning that the IE addressed or partially addressed at least two-thirds of the elements in the second tier (Table 1). Fourteen percent of the IE reports were of acceptable quality, meaning they addressed or partially addressed between half and two-thirds of the elements in the second tier. Thirty-seven percent of IE reports were of lower quality, having addressed or partially addressed fewer than half of the elements in the second tier. The remaining 46 percent of IE reports did not meet USAID's definition of an IE and were excluded from the full review because their findings could not be confidently attributed to the intervention evaluated. The remaining second-tier findings in this section correspond to the 72 IEs that met USAID's IE definition.

FIGURE 18: SECOND-TIER SCORE



Unlike the first-tier score, the association between second-tier scores and FY is small (Cramer's V = 0.2184), due to the modest increase over time in the inclusion of additional quality elements that strengthen the credibility of IE findings so they can be used to make decisions (Figure 19).

FIGURE 19: SECOND-TIER SCORES BY FISCAL YEAR (N=133)

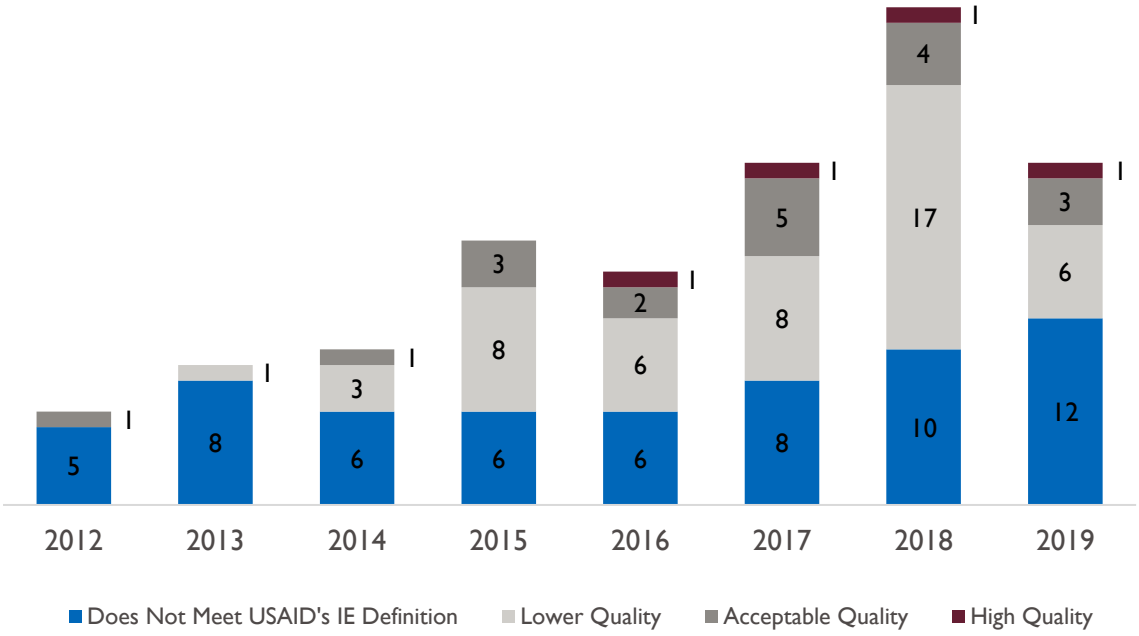
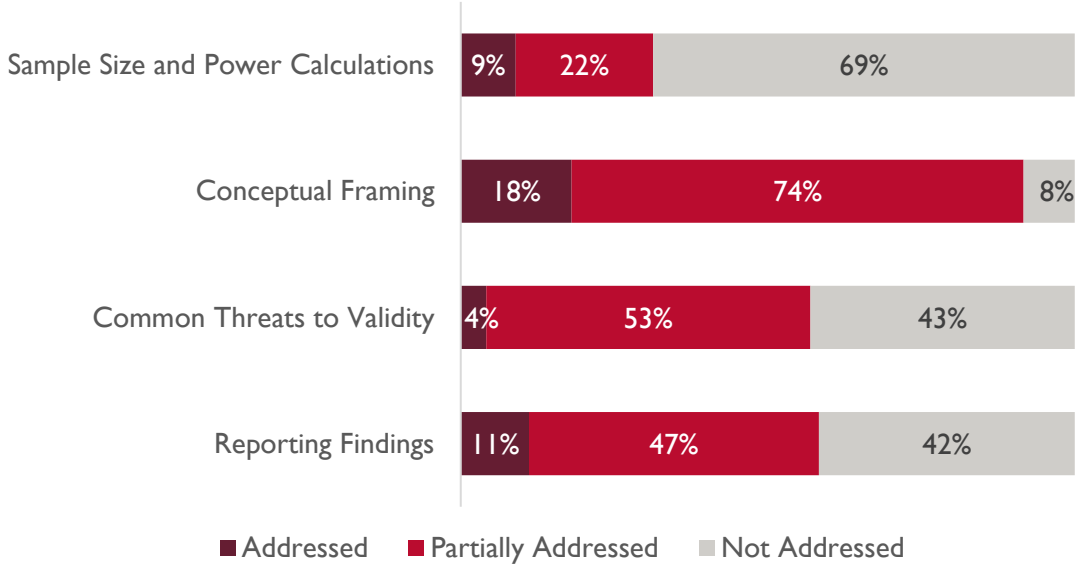


Figure 20 breaks down the second-tier score by domain. The team found that most IE reports did not address quality criteria at this level. Although most IE reports partially addressed quality criteria related to conceptual framing, the team found substantial limitations in addressing elements related to sample size calculations, common threats to validity, and reporting of findings.

FIGURE 20: SUMMARY OF SECOND TIER QUALITY ELEMENTS (N=72)



SAMPLE SIZE AND POWER CALCULATIONS

The second-tier score factors in three additional elements of power calculations: (1) accounting for expected take-up to determine sample size, (2) accounting for expected attrition to determine sample size, and (3) accounting for the intracluster correlation (ICC) in cluster-level designs to determine sample size.

These three elements are important to include in power calculations because most USAID IEs are effectiveness studies of real-world programs, where not everyone participates in the intervention when it is delivered (incomplete take-up) or units drop out during implementation or cannot be reached at follow-up (attrition). In addition, two-thirds of USAID's IEs have cluster-level designs, meaning that they assign treatment to intact groups or clusters (e.g., schools, villages) rather than individuals. This introduces another level of variability between individuals and within and between clusters, which must be taken into account by including an estimate of the ICC.

Absence of these elements in power calculations can lead to overly optimistic (too small) sample sizes and to severely underpowered studies. For example, the needed sample size is inversely proportional to the square of the take-up rate. Therefore, low take-up massively increases the sample needed to detect a desired effect. This is why to be powered to detect the same effect size with 50 percent take-up (i.e., only half the units offered the treatment actually show up/participate), the IE would need to increase sample size to four times more people than with 100 percent take-up.[14] Similarly, power calculations are very sensitive to the ICC estimate used, so excluding this factor also results in overly optimistic sample sizes.

The study found that only three IEs included expected take-up in the power calculations and 10 IEs included expected attrition rates (Figure 21). This does not mean that the rest of the IEs were underpowered[15] but rather that the reader cannot confirm that null effects are not due to the study being underpowered. Almost all IEs with cluster-level designs clearly defined the cluster level, but fewer than one-third (13 IEs) reported or mentioned the ICC (Figure 22). Of those 13 IEs, 10 explained the basis for the ICC selected (e.g., using other available datasets to calculate the ICC).

FIGURE 21: EXPECTED TAKE-UP AND ATTRITION INCLUDED IN POWER CALCULATIONS (N=68)

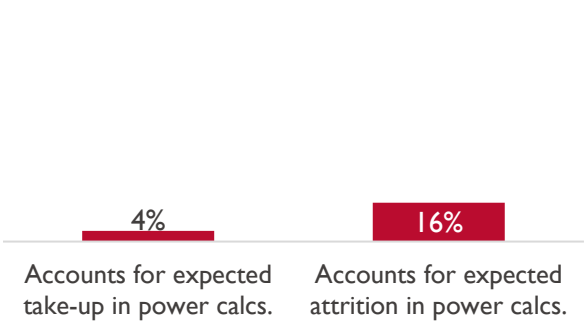
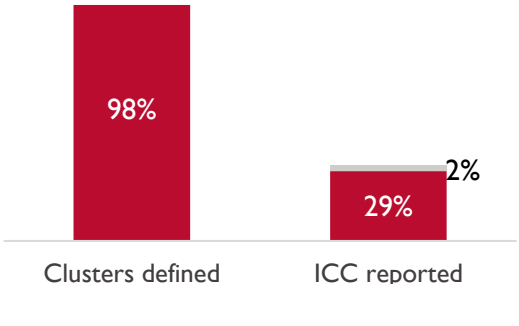


FIGURE 22: CLUSTERS DEFINED AND ICC REPORTED IN POWER CALCULATIONS (N=45)



Note: Figure 21 excludes the four IEs that relied on extensive administrative data as the primary data source. Figure 22 only includes the 45 cluster-level IEs.

[14] Development Impact blog post, "Power Calculations 101: Dealing with Incomplete Take-up" (McKenzie 2011).
[15] As with other quality elements, it is possible that IEs that did not include power calculations in the final report did conduct them previously and accounted for these factors. However, the reviewers did not search for design reports to confirm.

CONCEPTUAL FRAMING

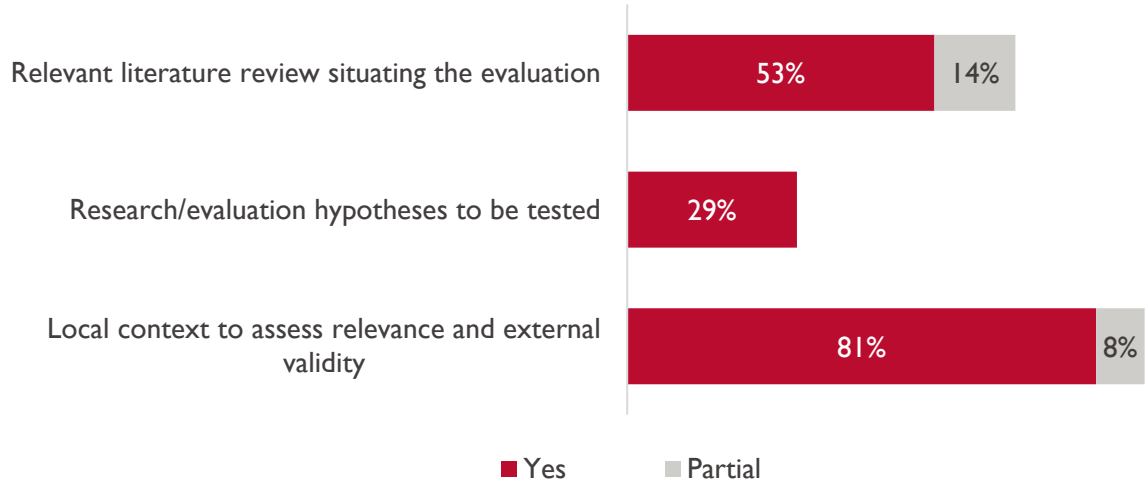
The second-tier score factors in three additional elements of conceptual framing: (1) a relevant literature review to situate the evaluation within the existing evidence and the relevant policy area, (2) research hypotheses to be tested, and (3) information on the local context to assess the evaluation's relevance and external validity.

A literature review of evidence and knowledge gaps helps explain the evaluation's theoretical framework and motivation and how the evidence generated from the evaluation will contribute to global knowledge. About half of the IE reports (53 percent) included a literature review (Figure 23). Some of the other IEs may have included a literature review in the design report but not in the final report; however, the reviewers did not search the design reports to confirm.

Supporting the theory of change with the specific hypotheses to be tested throughout the causal chain helps articulate the evaluation's aim and scope more strongly. The hypotheses specify the key outcomes of interest, the steps along the causal chain to be measured, and the subgroup analysis to be done. However, only 29 percent of IEs included the research hypotheses to be tested (Figure 23).

Information on the local context enables the reader to assess the evaluation's relevance and external validity. It can also help set realistic expectations for the results' value and scale and explain implementation risks. Eighty-one percent of IE reports included detailed information on the local context of the activity/intervention(s) being evaluated (Figure 23).

FIGURE 23: CONCEPTUAL FRAMING ELEMENTS (N=72)



COMMON THREATS TO VALIDITY

Although program implementation is mostly out of the evaluator's control, information on implementation fidelity and other common threats to evaluation validity is critical for the IE findings' credibility and use for decision making. In addition, fidelity and compliance are key assumptions underlying any theory of change, so information on these factors also helps explain the causal pathways and the reasons for changes (or lack of changes) in outcomes.

Implementation fidelity refers to whether treatment was delivered as planned, to the units assigned to receive treatment, and in the dosage and frequency intended. Sixty percent of the IE reports did not discuss implementation fidelity at all (Figure 24)—they did not provide any information about whether

implementation followed the intended plan. Although the absence of this discussion does not necessarily indicate issues with implementation fidelity, it leaves the reader without an understanding of what was actually done and how implementation could have affected the impact estimate. Fourteen percent of IE reports noted that treatment was delivered as planned, 18 percent that treatment was not delivered as planned due to minor changes, and 8 percent that treatment was not delivered as planned due to major changes (Figure 25).

The review also searched for discussion on four key challenges that are common threats to an evaluation's validity, including (1) take-up of treatment by units assigned to the treatment group; (2) contamination across groups, whereby units assigned to the control/comparison group received the intervention; (3) attrition through individuals dropping out of the program, inability to track individuals at the follow-up survey of a panel design, or removal of clusters from the evaluation; and (4) non-response or missing data in sampled units. The presence of these threats to validity is not a sign of low quality in itself as the threats can be addressed and accounted for statistically in the estimation of treatment effects (as long as they are not overly prevalent). For example, attrition may occur for various reasons but is problematic only if it correlates with observable characteristics so that the resulting follow-up sample is no longer representative of the population of interest or if there is differential attrition between the treatment and comparison group so that the resulting follow-up sample is no longer balanced between the groups. Therefore, the study reviewed whether the IEs discussed attrition and assessed how it affected the IE findings.

The study reviewed whether the IE reports discussed these four elements in detail, partially in vague terms, or not at all. As with implementation fidelity, if the IE report did not discuss these four elements, the team cannot distinguish between cases where "this was not an issue" and cases where "this was an issue but was not discussed." However, these are common challenges to conducting IEs, so a discussion of them in the IE report strengthens the explanation and credibility of the findings and serves as a proxy for quality. The study found that these common threats to validity were not typically discussed in the IE reports. Seventy percent did not discuss contamination across groups, 56 percent did not discuss non-response or missing data, and 50 percent did not discuss actual take-up (Figure 24). Attrition was more commonly discussed, with 39 percent reporting attrition rates and assessing how it affected the IE findings, 23 percent reporting it but not assessing how it affected the findings, and 38 percent not discussing it at all (Figure 28).

FIGURE 24: COMMON THREATS TO VALIDITY NOT DISCUSSED (N=72)

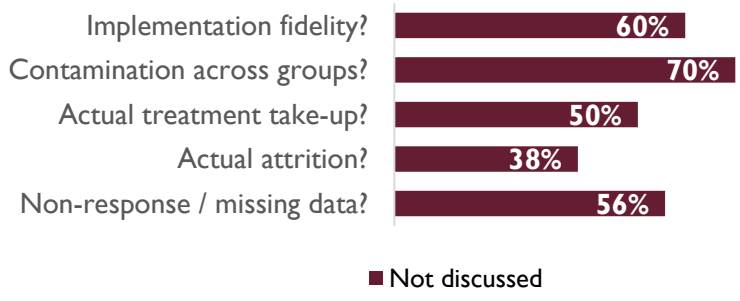


Figure 25 to Figure 29 provide more details on how IE reports discussed these common threats to validity.

FIGURE 25: IMPLEMENTATION FIDELITY

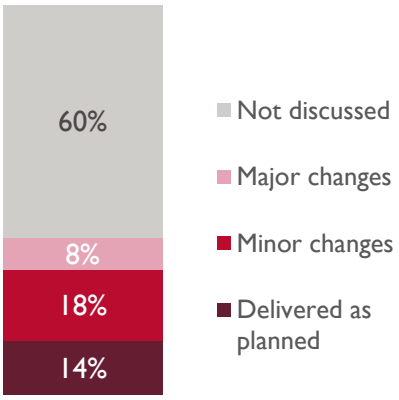


FIGURE 26: NON-COMPLIANCE

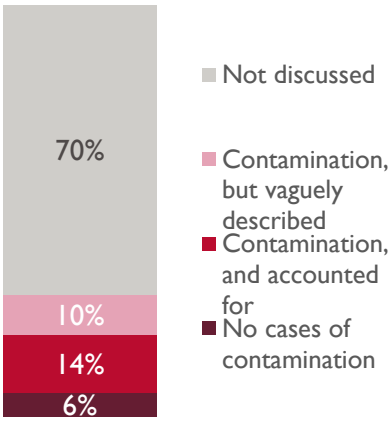


FIGURE 27: ACTUAL TREATMENT TAKE-UP

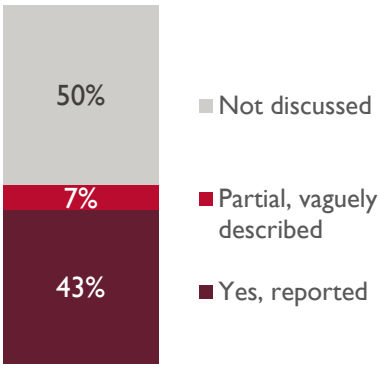


FIGURE 28: ACTUAL ATTRITION

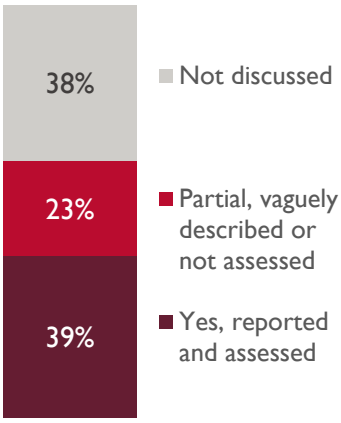
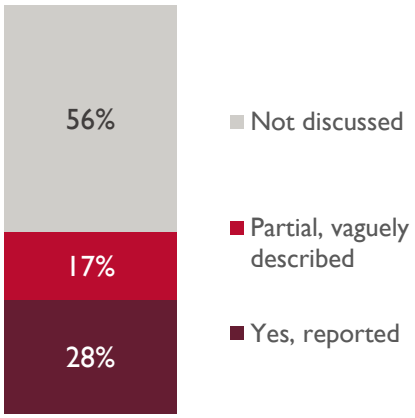


FIGURE 29: NON-RESPONSE/MISSING DATA



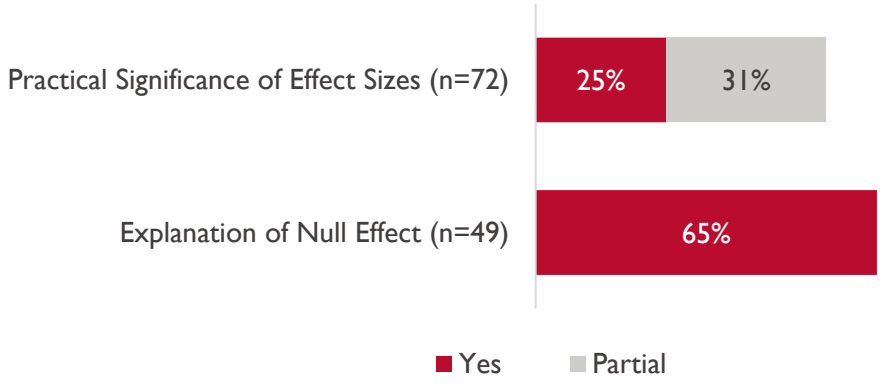
REPORTING FINDINGS

The second-tier score factors in two additional elements on reporting findings: (1) discussion of practical significance of impact estimates in terms of the effect size implications, and (2) discussion of why an effect was not detected.

Whereas statistical significance indicates whether an effect exists (i.e., the effect is not due to sampling error), practical significance refers to the magnitude of the effect and whether the effect is large enough to be of practical value. For example, a three percent increase in employment due to the program may be statistically significant, but its statistical significance says nothing about the size and potential meaning of this observed effect.[16] A discussion of practical significance can include comparison of the sample with a meaningful reference group or discussion of the impact estimate relative to the confidence intervals or as a standardized effect size. The study found that only 25 percent of IEs clearly explained practical significance and an additional 31 percent vaguely mentioned the effect's magnitude (e.g., "large effect") without describing what it was based on (Figure 30).

A discussion of null effects is also useful for readers to better understand the IE findings and use them appropriately. Lack of detected impact can be due to intervention ineffectiveness, limited power to detect changes, or even implementation issues. Forty-nine of the 72 IE reports reviewed (68 percent) did not detect an effect in at least one of their primary outcome measures. Of these 49 IE reports, 65 percent provided a potential explanation of why an effect was not detected.

FIGURE 30: MEANINGFUL EXPLANATIONS OF FINDINGS OR NULL EFFECTS



COST EFFECTIVENESS

An additional area of interest not factored into the quality scores in this study is the use of cost-effectiveness analysis to link an intervention's effectiveness to its costs. To influence resource allocation, USAID stakeholders need not only effectiveness evidence but also information about what it costs to deliver the intervention and how much impact can be achieved with a given dollar amount. Value for money analysis is particularly useful because resources are limited and trade-offs on resource allocation are regularly made. However, only 11 IEs of the 72 IE reports reviewed (15 percent) conducted cost-effectiveness analysis of the activity/intervention using the cost data and impact estimates measured.[17]

[16] The likelihood of statistical significance increases with larger sample sizes, even when differences between treatment and comparison groups are very small.
[17] An additional four IE reports provided cost data for either delivering the activity/intervention being evaluated or for the entire broader project. However, they did not provide a complete cost-effectiveness analysis.

CONCLUSIONS

Overall, the study findings show some quality improvements over time but critical gaps need to be addressed to improve IE quality at USAID. While the quality of IEs that meet USAID's IE definition (i.e., provided statistical justification for the comparison group's validity) has increased over time, 46 percent of USAID's IE reports did not meet USAID's IE definition. Eight years after USAID's evaluation policy was put in place, there continues to be two to four IE reports published each year without a comparison group and the number of IE reports that do not provide statistical justification for the comparison group has started to increase again in the past two years. In addition, almost all (93 percent) quasi-experimental IE reports with a non-equivalent groups design and 29 percent of IE reports with statistical matching or other statistical methods designs did not meet USAID's IE definition. Therefore, ensuring that IE reports, in particular those with non-equivalent groups design, provide statistical justification for the comparison group's validity is a critical gap that needs to be addressed to improve IE quality. This ensures the reader that the IE findings can be attributed to the intervention evaluated, but more importantly it provides USAID stakeholders with an evaluability checkpoint. If this step is done early on, USAID evaluation managers can review the comparison group's validity and decide on how to proceed with the evaluation.

The study findings also show that quality elements that are applicable to performance evaluations, such as descriptions of the local context, treatment and outcome measures, and data collection and analysis methods, were prevalent in the IE reports. However, quality elements that are not expected in performance evaluations, such as a literature review, theory of change, research hypotheses, power calculations, and common threats to validity, were not usually included in the IE reports. Since performance evaluations make up the majority of evaluations at USAID and most guidance on developing evaluation reports is not specific to IEs, it would make sense that quality elements that overlap with performance evaluations were more prevalent. Nonetheless, not all quality elements that are explicit criteria in USAID's reporting guidance and thus would be applicable to performance evaluations, such as listing the evaluation questions and describing the evaluation purpose, were prevalent in IEs. This implies a gap in guidance specific to IEs and presents an opportunity for USAID to reinforce its general evaluation reporting criteria as well.

The study findings also reveal the need for IE reports to shift from simply answering *whether* an activity/intervention is effective (e.g., only reporting impact estimates and statistical significance) to answering *why* there was an impact or lack thereof. The latter requires laying out a theory of change and defining outcome measures along the causal pathways, incorporating qualitative methods, conducting implementation fidelity monitoring to address common threats to validity, and providing explanations for null effects. It also requires building stronger relationships between evaluation teams and implementing partners so that there is consistent communication throughout the duration of the IE (not only at data collection points), collaboration to collect and share necessary monitoring data, and commitment to discuss implementation plans that may put at risk the IE's validity and any evaluation decisions that have implications for implementation plans. This shift in IE management and reporting will result in more credible findings that can be used by USAID stakeholders to make decisions and that contribute to the global knowledge base.

Given USAID's commitment to strengthen evidence-based development, IE findings need to be more easily accessible to incorporate into practice. The study shows the need for more discussion about the practical significance of impact estimates to interpret how large of an effect is the reported point estimate. This provides needed information for USAID stakeholders to decide whether the benefits are large enough to justify allocating resources in that activity/intervention. It also enables comparison to reported effects from other IEs.

Finally, the study assessed two factors that were not included in the quality scores but are relevant in USAID's commitment toward accountability: ethics considerations and cost effectiveness. First, ethics considerations are not required or recommended per USAID evaluation guidance; however ensuring confidentiality and preventing undue harm to evaluation participants is a standard practice for research involving human subjects. The study found that while all IEs gathered information from people, only 57 percent mentioned receiving informed consent for data collection, and 28 percent mentioned institutional review board approval. Second, data on cost effectiveness are an important complement to an IE and a critical component for decisionmakers to allocate limited resources efficiently. The study found that only 15 percent of IEs conducted a cost-effectiveness analysis. While low, this figure is similar to the estimated percent of IEs with any value for money analysis at the World Bank (19 percent) and in the International Initiative for Impact Evaluation's (3ie's) IE repository (14 percent).[18] These findings present an opportunity for USAID to include ethics considerations into evaluation guidance and reporting requirements and to operationalize the evaluation policy's call for cost-effectiveness.

In conclusion, while 47 percent of USAID IE reports are of acceptable or high quality with respect to basic quality elements (first tier), USAID must increase the share of IE reports that meet quality standards to generate credible findings that can be used to make decisions (second tier).

RECOMMENDATIONS

Based on the findings and conclusions presented above, the team recommends the following actions for USAID to improve IE report quality.[19]

1. **PPL/LER and USAID evaluation managers should reinforce that, per USAID's evaluation policy, IEs must include a comparison group.** Since 18 percent of USAID's IE reports, published at a rate of 2 to 4 IEs each FY, do not have a comparison group, it is crucial for this criterion to be followed to improve the quality of IEs at USAID. PPL/LER should emphasize this criterion with USAID evaluation managers and, in turn, USAID evaluation managers should use this criterion as a screen to decide on whether an impact evaluation is the most appropriate evaluation.
2. **PPL/LER should provide updated, detailed guidance on specific elements that should be included in final IE reports.** Although some of these elements may seem too technical for a general audience, they can be included as annexes so that the body of the IE report focuses more on discussing the findings and their implications. The new standard guidance should state a preference for including the following elements (either in the body of the report or in an annex):
    - **Statistical justification of the comparison group's validity.** This element appears usually in the form of a balance table with t-tests showing that the treatment and comparison groups were highly similar in key variables related to the treatment and outcome measures (demographics, behavior, outcomes themselves) before the intervention. It may also include additional analysis, such as an assessment of the parallel-trends assumption for multiple periods before the intervention or graphs to indicate the quality of a matched sample or the validity of a cut-off threshold for a regression discontinuity design. In addition to including this element in the final IE report, it should also be done and included

[18] Brown, Elizabeth, and Jeffery Tanner. "Integrating Value for Money and Impact Evaluations : Issues, Institutions, and Opportunities," World Bank Policy Research Working Paper. October 2019. No 9041, http://documents.worldbank.org/curated/en/862091571145787913/Integrating-Value-for-Money-and-Impact-Evaluations-Issues-Institutions-and-Opportunities.

[19] USAID released an update to ADS 201 on October 28, 2020, after this report had been drafted. The ADS 201 revisions align with some of the recommendations outlined in this report.

in the IE baseline report. This provides USAID stakeholders with an evaluability checkpoint to decide on how to proceed with the evaluation.

- **Explicit evaluation questions linked to the evaluation purpose.** Although reporting the evaluation questions is already an explicit criterion in USAID's reporting guidance, the IE-specific guidance should reinforce this criterion. This reinforcement would make it easier to ensure that IE reports adequately address all evaluation questions. In addition, the evaluation questions should be linked to the evaluation purpose, which must be explicit on the evaluation's intent—that is, the decisions and actions the evaluation is intended to inform.
- **A theoretical framework including a literature review, a theory of change, and specific hypotheses to be tested.** Specific guidance on these elements is particularly critical as these elements are not expected in Agency performance evaluations.
- **Defined and operationalized outcomes in the methodology section before presenting findings** (or referenced in the methodology section and included in an annex). Outcomes can be defined and measured in different ways. A clear operational definition of metrics will provide a better understanding of how the IE is measuring changes. It will also make it easier for USAID stakeholders to look across evaluations and aggregate learning that is relevant and comparable.
- **Specific power calculation parameters.** The key parameters that should be included in the final IE report are the power level and MDES, as well as the basis for the MDES used, and—for cluster-level designs—the ICC and basis for the estimate used. Power calculations should also adjust sample size to account for expected take-up and attrition estimates.
- **Detailed and complete information on common threats to validity.** IE reports should provide information on how the intervention was delivered and whether the implementation plan was followed. Fidelity and compliance are key assumptions of any theoretical framework and lack thereof can dilute impact or result in null effects.
- **Reporting findings that include point estimates and statistical significance as well as the control group mean** to interpret how large of an effect is the reported point estimate. This provides insights into the practical significance of the reported effect and enables comparison to reported effects from other IEs.
- **Discussion of null effects.** In cases where no impact is detected, IE reports should discuss potential explanations for the null results, such as intervention ineffectiveness, limited power to detect changes, or even implementation issues.
- **Actionable recommendations** that advise USAID decisionmakers on the implications of the IE's results for continuing/discontinuing, replicating, or scaling up the activity/intervention.

3. **PPL/LER should develop a standard IE report template and review checklist to minimize the omission of important quality elements.** The current USAID evaluation report and review checklist is useful but not specific to IE reports. The review instrument used in this study can be adapted into an IE report checklist tool and shared across the Agency. This is a low-cost way of increasing the value of USAID's investments in IEs. While it may require some additional efforts by USAID evaluation managers and evaluation teams to review IE reports against the checklist and to document more detailed information, the cost of doing so is low compared to funding an IE but not being able to use or have confidence in its results because of poor reporting.

4. **USAID evaluation managers should conduct evaluability assessments to ensure that only IEs that meet USAID's IE definition (i.e., adequately powered study to measure changes with a valid comparison group) are funded.** Evaluability assessments provide decision points for USAID evaluation managers to determine whether, beyond meeting an Agency

requirement, there are valid reasons for undertaking an IE of a specific activity/intervention, and no overriding reasons for not doing so. These assessments are generally conducted prior to the decision to commission an IE, but can also be revisited throughout the design and implementation of an IE. USAID evaluation managers should proceed with funding only adequately powered IEs with valid comparison/control groups. A useful mechanism for addressing evaluability feasibility issues early on is to conduct an IE design workshop between the IE team and the activity implementing partners soon after USAID awards the activity.

5. **USAID evaluation managers should commission external peer reviews to assess the quality of evaluation designs and final reports, especially when there are gaps in internal technical capacity to adequately do so.** The additional reporting requirements and evaluability assessments proposed may require technical expertise that is not always available at individual missions or operating units. External peer reviews of evaluation designs ensure that plans for executing the IE are free from serious framing flaws and other technical impediments to the IE's success. Peer reviews of the final IE report can include a review of the proposed report checklist as well as a questionnaire to guide the assessment. Several USAID IEs have conducted external peer reviews but USAID does not currently have a formal process or protocol for conducting such reviews of IE designs.

6. **USAID evaluation managers should integrate implementation fidelity monitoring into IE SOWs.** Accounting for common threats to validity is an important quality element for IEs. However, gathering this information requires early coordination with implementing partners and resources. Including implementation fidelity monitoring into IE SOWs ensures that the adequate budget, timeline, and planning are allocated to gather this information and include it in IE reports.

7. **PPL/LER should provide guidance to shift IEs toward reporting more information to disentangle and explain effects.** USAID should provide guidance on how IEs can be prioritized for programs where specific interventions can be tested or how an IE design can be modified to account for or measure variance in interventions or in their delivery. In addition, USAID should promote greater use of qualitative methods to disentangle effects—only half of IEs with bundled interventions used qualitative methods to better understand causal pathways.

8. **USAID should integrate ethics considerations as an IE standard to align with its Scientific Research Policy.** The Agency's Scientific Research Policy outlines that USAID-funded research/evaluations must conform to legal and other requirements governing research with human subjects in the country where it is conducted. A study must be reviewed and approved or deemed exempt by a U.S.-based IRB. However, existing evaluation guidance does not integrate these ethics considerations. Aligning these two policies would close the gap in reporting and ensure USAID-funded IEs meet ethical standards of accountability and social responsibility.

9. **USAID should integrate the evaluation policy's call for cost effectiveness as an IE standard.** USAID's operational policy suggests that staff might want to analyze cost effectiveness as part of program development and design and that evaluations that are expected to influence resource allocation should include information on the intervention's cost structure and scalability, as well as its effectiveness. However, these suggestions are not requirements and are infrequently implemented. USAID should provide institutional support for cost-effectiveness analyses and guidance on when and how to conduct them. Instituting cost-effectiveness standards will place USAID as a thought leader in this space since value for money is still largely absent from most IEs at other agencies and donors.

ANNEX A: ACTIVITY STATEMENT OF WORK

Statement of Work: Impact Evaluation Quality Review for the Discussion Paper on Enhancing E3 Impact Evaluation Investments

This statement of work (SOW) is for a review of impact evaluation quality that was initiated as part of a Discussion Paper on impact evaluations related to sectors supported by USAID's Bureau for Economic Growth, Education, and Environment (E3). The Discussion Paper, Opportunities for Enhancing Returns on E3 Bureau Investments in Impact Evaluation, was developed by the E3 Analytics and Evaluation Project and finalized in August 2019. The Discussion Paper describes the emergence of impact evaluation in the development assistance community; provides the results of a preliminary review of quality characteristics of impact evaluations in E3 sectors in relation to other USAID impact evaluations produced between 2012 and 2016; and catalogues challenges to high-quality impact evaluations as well as promising approaches and tools for addressing those challenges, some of which had emerged from E3 technical offices.

1. Introduction and Background

Under this activity, the E3 Analytics and Evaluation Project will develop a companion report to the Discussion Paper that the E3 Analytics and Evaluation Project developed collaboratively with the E3 Bureau in 2016-2017, examining the quality of recent USAID impact evaluations. The Discussion Paper was based on the Project team's experience with the planning and implementation of impact evaluations it supported. Initially developed as a "White Paper" on impact evaluation findings and lessons, the Paper was later split into two documents during consultations in September 2017 between the Project team, the E3 Bureau, and the Bureau for Policy, Planning and Learning's Office of Learning, Evaluation and Research (PPL/LER). Those two documents are (1) the Discussion Paper on challenges to conducting impact evaluations and promising solutions, and (2) the forthcoming paper reviewing the quality of a set of USAID impact evaluations (which is the focus of this SOW). Annex 1 of this SOW provides a detailed summary of this historical process.

2. Activity Purpose, Audience, and Intended Use

This activity will enhance the E3 Bureau's ongoing efforts to encourage and support the design, implementation, and utilization of rigorous impact evaluations in E3 sectors and on cross-cutting issues on which the Bureau has a leadership role. The activity will also produce a separate summary of action recommendations for the E3 Bureau stemming from the findings documented in the two reports.

Purpose

Impact evaluations are an important avenue through which the E3 Bureau contributes evidence for decision-making on foreign assistance policies, strategies, projects, and activities. This is consistent with the Agency's evaluation policy, Automated Directives System (ADS) 201 guidance, and Goal 4.1.1. of the Department of State – USAID Joint Strategic Plan, 2018-2022, which states, "By 2022, increase the use of evidence to inform budget, program planning and design, and management decisions."

The purpose of this activity is to enhance impact evaluation practice in E3 sectors. Within the E3 Bureau, the Office of Planning, Learning, and Coordination (PLC) and office-level monitoring and evaluation (M&E) leads share responsibility for encouraging, supporting, and monitoring the quality, dissemination, and follow-up on impact evaluations conducted in E3 sectors and on cross-cutting issues for which the Bureau has oversight (including E3-funded evaluations as well as mission-funded evaluations about sectors and topics the Bureau supports). Products from this activity will provide E3

staff with evidence and tools to ensure impact evaluation practice in E3 sectors is consistent with Agency policy and consciously aspires to professional norms for impact evaluation quality.[20]

Audience

The primary audiences for this activity are E3/PLC staff including M&E leads in E3 offices as well as their counterparts in missions who focus on evaluations in E3 sectors. Additional audiences for products from this activity include PPL/LER and M&E leads in other USAID/Washington bureaus and missions that undertake impact evaluations. Secondary audiences include other USAID staff with an interest in impact evaluations and implementing partners involved in planning, conducting, and disseminating impact evaluations.

Intended Uses

The report on impact evaluation quality is intended to be used by E3/PLC and other Agency evaluation advisors to pinpoint consistent impact evaluation weaknesses and provide assistance, guidance, or training as needed to improve the quality of impact evaluations in which their operating units invest. It is expected that the Discussion Paper will be used by those involved with impact evaluations to anticipate and effectively address challenges that can undermine such evaluations.

3. Gender Considerations

It is not expected that this activity will encounter gender-specific issues but if it should, they will be reported and addressed by activity products as appropriate.

4. Existing Information Sources and Cross-Activity Collaboration

This activity will benefit from previous work on the original 2017 draft Discussion Paper. It will also benefit from additional impact evaluations completed by USAID and posted on the Development Experience Clearinghouse (DEC) between late 2016 and late 2018, as well as additional interviews undertaken with M&E leads in other USAID/Washington bureaus at the end of 2018.

5. Activity Methods

Background on the Discussion Paper on Challenges and Promising Solutions for Impact Evaluations

The Opportunities for Enhancing Returns on E3 Bureau Investments in Impact Evaluation Discussion Paper was drafted in June 2017 and revised in November 2017 by the E3 Analytics and Evaluation Project. USAID staff then provided suggestions for improving the draft Discussion Paper including obtaining inputs from M&E leads in bureaus other than E3 to capture additional experience with challenges, impact evaluation utilization, and promising solutions. Reviewers also suggested that the Paper be reconfigured to better foster the application of promising solutions through more direct and user-friendly connections between challenges and potential solutions. The Project team's analysis also highlighted the need for greater distinction between existing USAID information and approaches for addressing challenges already documented and disseminated through USAID's Evaluation Toolkit and innovative practices emerging from experience and staff initiatives in E3 and other bureaus. Methods the team applied to finalize the Discussion Paper included additional interviews with staff in USAID bureaus

[20] Professional norms for impact evaluation quality include those set forth in USAID's 2013 Technical Note: Impact Evaluations as well as other established standards for rigorous impact evaluations such as those governing entry into the U.S. Department of Education's "What Works" Clearinghouse, acceptance for the International Initiative for Impact Evaluation (3ie) evaluation hub repository, and other sources described in the draft quality rating instrument in Annex 2.

other than E3 and reordering the report's contents in response to suggestions from USAID reviewers. The Project team also analyzed additional information obtained since 2017 to expand the examples and approaches included in the Paper. Following these activities, the Project team submitted the final Discussion Paper for USAID's approval in August 2019.

Preparing the Report on Impact Evaluation Quality

The impact evaluation quality review report will reprise and advance the section of the 2017 draft White Paper that addressed this topic. In addition to describing the evolution of impact evaluation in the international development community, that section reported on several key quality characteristics of 24 verified USAID impact evaluations available on the DEC published between 2012 and 2016.[21] The characteristics addressed went beyond those included in previous USAID-wide and E3 Bureau meta-evaluations that examined quality in terms of evaluation report compliance with USAID evaluation policy and guidance. For example, the White Paper's initial review of impact evaluation quality included factors such as the type of evaluation design utilized; whether the sample size was adequate based on a power calculation; and whether the report described impact in effect size terms. All these quality features are suggested in USAID's 2013 Technical Note: Impact Evaluations, but are not expected in Agency performance evaluations.

The impact evaluation review instrument to be used in finalizing this aspect of the work begun under the White Paper will more comprehensively test USAID impact evaluations in E3 and other sectors against professional impact evaluation norms. It will also highlight impact evaluation design and implementation choices evident in reports, such patterns with respect to power levels on which samples are based, attrition experience, effect sizes reported, and the types of evaluation questions completed USAID impact evaluations most frequently address. The instrument to be used is designed to be as objective as possible based on standards rather than the subjective judgments of expert evaluators. This approach is consistent with USAID's basic evaluation report review checklist in its Evaluation Toolkit. The intent with a standards-based tool is for USAID staff and evaluation teams to be able to learn and apply the checklist criteria as they carry out impact evaluations. Annex 2 provides the draft updated instrument. Under this activity it will be used to update reviews of the 24 impact evaluations rated in the draft White Paper as well as approximately additional impact evaluations on the DEC published from 2017 through FY2019.

Methods to be involved in carrying out the impact evaluation quality review include updating the review instrument based on advances made in preparation for presentations Project team members made at the American Evaluation Association in November 2018 with E3/PLC and PPL/LER staff and feedback from those presentations; updating impact evaluation review results for evaluations completed between 2012-2016 that the team originally rated in 2017; reviewing additional USAID evaluations completed in 2017, 2018, and FY2019; and updating the team's analysis and findings on verified impact evaluation quality based on the above.

In addition to rating impact evaluations, the Project team will work with E3 and PPL/LER staff to efficiently identify stories along the evidence-to-action chain that demonstrate how these studies have affected policy and program cycle decisions in USAID and by its partners. Including even a small effort to understand use reflects USAID's Evaluation Policy assertion that the success of the Agency's impact

[21] "Verified" refers to the screening process the Project team used to determine which of the 43 evaluation reports at that time that included "impact" in their title met USAID's minimum standard as described in its Evaluation Policy and ADS 201 to "measure the change in a development outcome that is attributable to a defined intervention. Impact evaluations are based on models of cause and effect and require a credible and rigorously defined counterfactual to control for factors other than the intervention that might account for the observed change."

evaluations "will not be predicated on the number of evaluations done, or stored within a database, or even solely upon the quality of the findings. Evaluation is useful only insofar as it provides evidence to inform real-world decision making…We'll be successful if and when the evaluation work of USAID contributes to greater development effectiveness."

The analysis plan for this impact evaluation quality review report includes calculating the frequency with which USAID impact evaluations display features that are expected or encouraged in impact evaluations as well as examining differences in performance on these characteristics based on the evaluation design employed, year completed, and sector, to the degree those features are associated with differences on the characteristics the review examines.

Figure 1 displays the preliminary number of impact evaluation reports published from 2012 to 2018 that this activity will review. The Project team will finalize the list of impact evaluations reports in consultation with E3 and PPL/LER staff. As the figure indicates, the number has mostly increased over time, which is indicative of the numbers of impact evaluations initiated in the first few years after USAID's evaluation policy requiring impact evaluations for pilot activities and innovative interventions. Given that many impact evaluations involve baseline and endline comparisons spanning three to five years, the increase in reports published in 2017 is suggestive of the number initiated between 2012 and 2014. The same time span considerations make it likely that impact evaluation reports posted in 2014 and 2015 were initiated before the evaluation policy requirement for this type of evaluation.

Table 1 divides the number of impact evaluations published into three periods: those verified for 2012-2016 (which the draft 2017 White Paper examined and this review will update) and 2017-2018 (which will also be included in this review), as well as an estimated but not yet verified count for FY2019. The number of evaluations to be examined under this activity is roughly the same for 2012-2016 and 2017-2018. The design of these evaluations is split about the same in both periods, with roughly one-third using experimental designs and two-thirds being quasi-experimental designs.

Table 1: Verified Impact Evaluations by Type of Design and Time Period

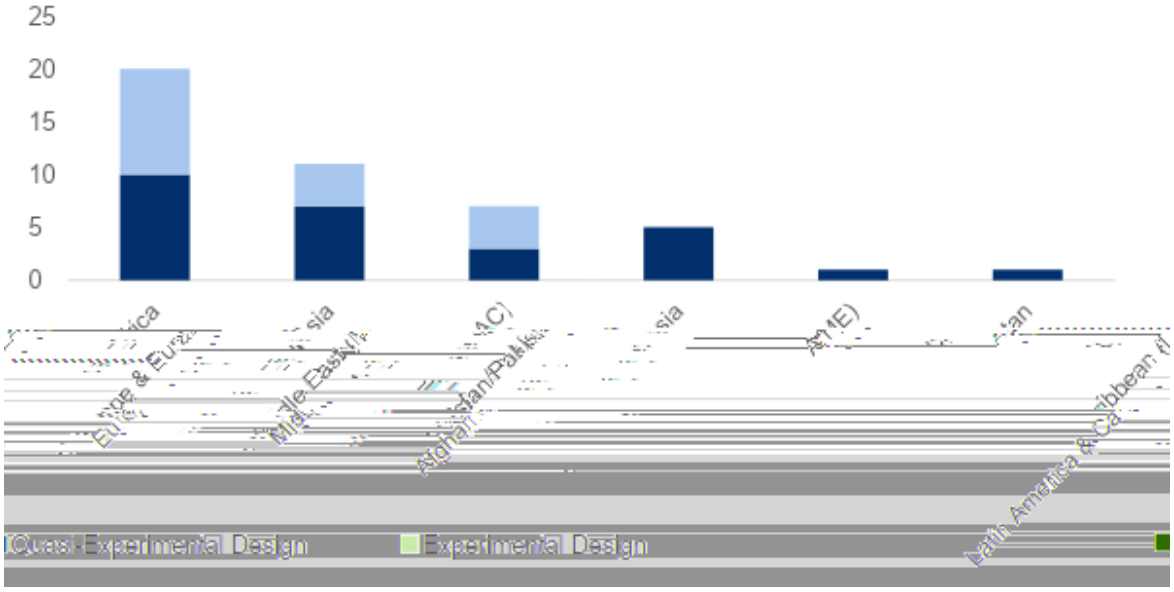
Design Type	2012-2016	2017-2018	FY2019 (Estimated)	Total
Experimental Design	8	9	4	**21**
Quasi-Experimental Design	16	12	5	**33**
Total	**24**	**21**	**9**	**54**

As Table 2 shows, E3 sectors were the most frequent topic covered by these 45 verified impact evaluations from 2012-2018 (40 percent). Figure 2 shows the same set of impact evaluations distributed by regions in which USAID works and the type of impact evaluation conducted.

Table 2: Distribution of Verified Impact Evaluations by Supporting Technical Bureau (2012-2018)

USAID Bureau	# of Verified Impact Evaluations in the DEC (2012-2018)	Percentage	Notes on Distribution
E3	18	40%	• 10 on education • 1 on gender
DCHA	16	36%	• 4 on youth
Global Health	6	13%	
BFS	5	11%	

Figure 2: Distribution of USAID Impact Evaluations by Region and Type (2012-2018)



6. Strengths and Limitations

With respect to the Discussion Paper on challenges to high-quality impact evaluations and promising solutions, the range of approaches for addressing challenges is limited by the range of USAID representatives from whom experiences and approaches were obtained. Most ideas from the E3 Bureau about how to address challenges came from offices that have undertaken multiple impact evaluations. The Project team also obtained input from USAID staff in DCHA, BFS, the U.S. Global Development Lab, and the Latin America and the Caribbean Regional Bureaus, but not from the Global Health Bureau, other regional bureaus or missions. The team sought additional ideas and examples through direct contacts in USAID/Washington and PPL/LER, following the presentation it delivered to PPL/LER's Evaluation Interest Group around the November 2017 draft Discussion Paper.

With respect to the report on impact evaluation quality, while the coverage of impact evaluations is complete and will allow comparisons based on type of impact evaluation design, year completed, and

bureau, the number of impact evaluation characteristics examined will be limited to the amount of time available for updating and adding new impact evaluation reviews. To date, the Project team's approach in selecting characteristics to examine has been to prioritize those most important for determining the quality and conclusiveness of impact evaluation results. That approach will be applied as well to the revised review instrument.

7. Deliverables and Reporting Requirements

The following deliverables are envisioned as part of this activity. Due dates are estimates and may be amended with concurrence from the USAID Contracting Officer's Representative.

Deliverable	Estimated Delivery Date
1. Pre-Tested and Finalized Impact Evaluation Review Instrument	o/a 3 weeks following USAID approval of this SOW
2. Draft Impact Evaluation Review Report	o/a 12 weeks following USAID approval of the finalized Review Instrument
3. Final Impact Evaluation Review Report	o/a 3 weeks after receipt of all written USAID comments on the Draft Review Report
4. Draft Action Recommendations	o/a 2 weeks after receipt of all written USAID comments on the Final Review Report
5. Final Action Recommendations	o/a one week after receipt of USAID comments on the draft Action Recommendations

All documents will be provided electronically to USAID no later than the dates approved by USAID.

The format of the two final reports will be agreed upon between USAID and the team. The format for the action recommendations submission that will accompany these reports will be agreed upon with USAID in advance.

8. Team Composition

USAID anticipates the following composition for the activity team.

Impact Evaluation Specialist – Irene Velez: Ms. Velez, an impact evaluation specialist and contributing author of the Discussion Paper, will oversee the execution of this study including the preparation of the reports and action recommendations.

Researchers: One to two evaluation specialists will support the impact evaluation specialist in completing the impact evaluation reviews using the updated instrument and will contribute suggestions for the action recommendations based on findings from their work under this activity.

Home Office support by the E3 Analytics and Evaluation Project will be provided as needed, including technical guidance, research assistance, administrative oversight, data analysis, and logistical support.

9. USAID Participation

An interactive and collaborative process is envisioned between the activity team and E3/PLC to carry out this activity. USAID will have input into the activity design through a review of draft product templates and the review instrument. The Project team will work closely with the USAID Contracting Officer's Representative to identify options for disseminating activity products.

10. Schedule

The table below provides a tentative timeframe for the main tasks under this activity. This schedule and the deliverable dates above will be finalized in coordination with the USAID Contracting Officer's Representative.

Tasks	Weeks Following SOW Approval																											
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28
Impact Evaluations Quality Review																												
Address USAID comments on draft instrument, pretest, and finalize it																												
Train researcher(s) and rate impact evaluations																												
Data analysis and draft report submission																												
USAID review																												
Report revisions and final submission																												
Action Recommendations																												
Development of action recommendations and submission																												
USAID review of draft																												
Action recommendations revisions and final submission																												

11. Estimated LOE and Budget

The team responding to this SOW will deliver a detailed estimated budget for USAID's review and approval prior to commencing implementation of the activity.

ANNEX B: FULL LIST OF IE REPORTS (N=133)

#	FY	Report Title	DEC URL	Country	Technical Area
1	2012	The VOICE impact evaluation in the Democratic Republic of the Congo : final report	https://pdf.usaid.gov/pdf_docs/pdacu337.pdf	Congo DR	Democracy & Governance
2	2012	Exploring the impact of the community based care for orphans and vulnerable children (CBCO) program	https://pdf.usaid.gov/pdf_docs/PDACT939.pdf	Kenya	Economic Growth
3	2012	RLS-I impact evaluation report, July 2012 : rule of law stabilization program -- informal component	https://pdf.usaid.gov/pdf_docs/pdacw029.pdf	Afghanistan	Democracy & Governance
4	2012	Programmatic impact evaluation in the energy sector in Kosovo : final report	https://pdf.usaid.gov/pdf_docs/pdacu492.pdf	Kosovo	Environment
5	2012	Preliminary evaluation : impact on health status of orphans and vulnerable children in Namibia of mobile primary care clinics provided by the Mister Sister Public Private Partnership	https://pdf.usaid.gov/pdf_docs/PNADZ328.pdf	Namibia	Health
6	2012	Final evaluation of ProAgro 2006-2012 : lessons learned from six years of cooperative agriculture development in Angola	https://pdf.usaid.gov/pdf_docs/PDACU683.pdf	Angola	Agriculture & Land
7	2013	Action against malnutrition through agriculture : Nepal child survival project, Kailali and Baitadi Districts, Far Western Region Bajura Expansion District : final evaluation report	https://pdf.usaid.gov/pdf_docs/PA00KMDV.pdf	Nepal	Health
8	2013	USAID/India innovations in family planning services : final evaluation report	https://pdf.usaid.gov/pdf_docs/PA00JQ4B.pdf	India	Health
9	2013	Evaluation of LAC higher education scholarships program	https://pdf.usaid.gov/pdf_docs/PDACX232.pdf	Mexico, Honduras, Guatemala	Education
10	2013	Evaluation : USAID/Dominican Republic Batey community development project	https://pdf.usaid.gov/pdf_docs/PDACY353.pdf	Dominican Republic	Economic Growth
11	2013	Quality and Humanization of Care Assessment (QHCA) A Study of the Quality of Maternal and Newborn Care Delivered in Mozambique's Model Maternities	https://pdf.usaid.gov/pdf_docs/PA00JS61.pdf	Mozambique	Health
12	2013	Bangladesh smiling sun franchise program (BSSFP) impact evaluation report	https://pdf.usaid.gov/pdf_docs/pdacu705.pdf	Bangladesh	Health
13	2013	Constituency dialogues and citizen engagement in Cambodia : findings from a mixed methods impact evaluation	https://pdf.usaid.gov/pdf_docs/pa00jqn9.pdf	Cambodia	Democracy & Governance

#	FY	Report Title	DEC URL	Country	Technical Area
14	2013	Agricultural credit enhancement program : midterm impact evaluation	https://pdf.usaid.gov/pdf_docs/pa00k2kg.pdf	Afghanistan	Agriculture & Land
15	2013	Golos impact evaluation : final report	https://pdf.usaid.gov/pdf_docs/pbaae653.pdf	Russia	Democracy & Governance
16	2014	Africa trade hubs export promotion evaluation	https://pdf.usaid.gov/pdf_docs/PDACX958.pdf	Sub-Saharan Africa	Economic Growth
17	2014	Education data for decision making (EdData II): National early grade literacy and numeracy survey - Jordan Intervention impact analysis report	https://pdf.usaid.gov/pdf_docs/PA00KH3M.pdf	Jordan	Education
18	2014	Tackling informality through tax reform IE -Georgia (IFC/Impact Program)	Non-DEC link	Georgia	Economic Growth
19	2014	Quasi-experimental assessment of the poverty reduction and alleviation project (2009-2012) : final report	https://pdf.usaid.gov/pdf_docs/PA00K63X.pdf	Peru	Economic Growth
20	2014	Primary math and reading (PRIMR) initiative : endline impact evaluation	https://pdf.usaid.gov/pdf_docs/pa00k27s.pdf	Kenya	Education
21	2014	Impact evaluation of the mayer hashi program of long-acting and permanent methods of contraception in Bangladesh	https://pdf.usaid.gov/pdf_docs/pa00k269.pdf	Bangladesh	Health
22	2014	Yes youth can! impact evaluation : final report	https://pdf.usaid.gov/pdf_docs/pa00jzqx.pdf	Kenya	Democracy & Governance
23	2014	Impact evaluation of supporting traditional leaders and local structures to mitigate community-level conflict in Zimbabwe : final report	https://pdf.usaid.gov/pdf_docs/pa00k5r5.pdf	Zimbabwe	Democracy & Governance
24	2014	Smallholder oil palm support (SHOPS) final impact evaluation	https://pdf.usaid.gov/pdf_docs/pa00k1k9.pdf	Liberia	Agriculture & Land
25	2014	Evaluation of the Impact of Malaria Control Interventions on All Cause Mortality in Children Under-five in Senegal	Non-DEC link	Senegal	Health
26	2015	Impact evaluation of crime prevention programs in Ciudad Juarez, Monterrey, and Tijuana	https://pdf.usaid.gov/pdf_docs/pa00m1t8.pdf	Mexico	Democracy & Governance
27	2015	Quantitative impact evaluation of the SHOUHARDO II project in Bangladesh	https://pdf.usaid.gov/pdf_docs/pa00kfcd.pdf	Bangladesh	Agriculture & Land

#	FY	Report Title	DEC URL	Country	Technical Area
28	2015	Improving health service delivery through community monitoring and non-financial awards : report to USAID	https://pdf.usaid.gov/pdf_docs/pa00mdrt.pdf	Sierra Leone	Health
29	2015	Retrospective impact evaluation of alternative development program in Huanuco, San Martin and Ucayali (2007-2012)	https://pdf.usaid.gov/pdf_docs/pa00kcbv.pdf	Peru	Economic Growth
30	2015	Impact evaluation of USAID/Indonesia's Kinerja program	https://pdf.usaid.gov/pdf_docs/pa00krbn.pdf	Indonesia	Democracy & Governance
31	2015	Strengthening and Evaluating the Preventing Malnutrition in Children under 2 Years of Age Approach – Burundi Follow-up Report: Children 0–23 Months	https://pdf.usaid.gov/pdf_docs/PA00K7C4.pdf	Burundi	Health
32	2015	Impact evaluation of USAID's community-based crime and violence prevention approach in Central America : regional report for El Salvador, Guatemala, Honduras and Panama	https://pdf.usaid.gov/pdf_docs/pbaab431.pdf	Central America	Democracy & Governance
33	2015	d.light solar home system impact evaluation : affordable access to energy for all : innovative financing for solar systems	https://pdf.usaid.gov/pdf_docs/pa00mdss.pdf	Uganda	Environment
34	2015	Improving agricultural competitiveness in Bosnia and Herzegovina : impact evaluation of USAID and Sida fostering agricultural market activity (FARMA)	https://pdf.usaid.gov/pdf_docs/pa00kf58.pdf	Bosnia and Herzegovina	Agriculture & Land
35	2015	Networks and information : an impact evaluation of efforts to increase political participation in Mozambique	https://pdf.usaid.gov/pdf_docs/pa00kq23.pdf	Mozambique	Democracy & Governance
36	2015	Using Learning Camps to Improve Basic Learning Outcomes of Primary School Children : Grant No. AID-OAA-F-13-00023	https://pdf.usaid.gov/pdf_docs/pa00kqjc.pdf	India	Education
37	2015	Impact evaluation of peace through development II (P-DEV II) : radio programming in Chad and Niger	https://pdf.usaid.gov/pdf_docs/pa00ktf3.pdf	Chad, Niger	Democracy & Governance
38	2015	Do early warning systems and student engagement activities reduce dropout? : findings from the school dropout prevention pilot program impact evaluation in Timor-Leste. Volume I : main findings	https://pdf.usaid.gov/pdf_docs/pbaad881.pdf	Timor-Leste	Education
39	2015	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	https://pdf.usaid.gov/pdf_docs/pa00k6z6.pdf	Ivory Coast	Health

#	FY	Report Title	DEC URL	Country	Technical Area
40	2015	Food security for flood-affected populations in Odisha : project surakhya : final impact evaluation report	https://pdf.usaid.gov/pdf_docs/pa00kp3c.pdf	India	Agriculture & Land
41	2015	Evaluation of the Impact of Malaria Control Interventions on All Cause Mortality in Children under-five in Ethiopia	Non-DEC link	Ethiopia	Health
42	2015	Global UGRAD Educational Exchange Impact Evaluation	Non-DEC Link	Pakistan	Education
43	2016	Impact evaluation : results-based financing in the Democratic Republic of Congo	https://pdf.usaid.gov/pdf_docs/pa00m2j7.pdf	Congo DR	Health
44	2016	Impact evaluation of the Georgia new economic opportunities (NEO) project : report on the endline impact evaluation of NEO's component 1, 2 and 3 activities	https://pdf.usaid.gov/pdf_docs/pa00kxkt.pdf	Georgia	Economic Growth
45	2016	Nigeria Reading and Access Research Activity Results of an approach to improve early grade reading in Hausa in Bauchi and Sokoto States	https://pdf.usaid.gov/pdf_docs/PA00KVM1.pdf	Nigeria	Education
46	2016	Agricultural Extension, Technology Adoption and Information Spillovers: Evidence from a Cluster Randomized Experiment	https://pdf.usaid.gov/pdf_docs/PA00SRVB.pdf	India	Agriculture & Land
47	2016	Ethiopia strengthening land tenure and administration program endline report : an impact evaluation of the effects of second-level land certification relative to first-level certification	https://pdf.usaid.gov/pdf_docs/pa00m3zp.pdf	Ethiopia	Agriculture & Land
48	2016	Supporting small and medium-sized enterprises in Bosnia and Herzegovina : impact evaluation of USAID and SIDA fostering interventions for rapid market advancement (FIRMA)	https://pdf.usaid.gov/pdf_docs/pa00kssc.pdf	Bosnia and Herzegovina	Economic Growth
49	2016	Providing free pregnancy test kits to community health workers increases distribution of contraceptives: results from an impact evaluation in Madagascar	Non-DEC link	Madagascar	Health
50	2016	Midterm impact evaluation of the Afro-Colombian and indigenous program (ACIP) 2011-2016	https://pdf.usaid.gov/pdf_docs/pa00mck7.pdf	Colombia	Democracy & Governance
51	2016	Impact evaluation of the USAID/aprender a ler project in Mozambique : final report	https://pdf.usaid.gov/pdf_docs/pa00m5d4.pdf	Mozambique	Education

#	FY	Report Title	DEC URL	Country	Technical Area
52	2016	Evaluating the effect of gender-equity maternal and child health programs In Uganda	https://pdf.usaid.gov/pdf_docs/PA00MNRJ.pdf	Uganda	Health
53	2016	USAID/Georgia external impact evaluation of the Georgian primary education (G-PRIED) project : endline impact evaluation report	https://pdf.usaid.gov/pdf_docs/pa00m256.pdf	Georgia	Education
54	2016	Impact evaluation of Feed the Future/USAID-ACCESO : agriculture and nutrition activities in Western Honduras from 2012-2015	https://pdf.usaid.gov/pdf_docs/pa00tcjk.pdf	Honduras	Agriculture & Land
55	2016	Final evaluation and scaling report : head safe, helmet on	https://pdf.usaid.gov/pdf_docs/PA00MJPT.pdf	Cambodia	Education
56	2016	KickStart – Washington State University (WSU) Research Study Milestone 11: Final Impact Evaluation Assessment Report	https://pdf.usaid.gov/pdf_docs/PA00SRV5.pdf	Kenya	Health
57	2016	Evaluation of the impact of malaria control interventions on all-cause mortality in children under-five in Mozambique	https://pdf.usaid.gov/pdf_docs/PA00TFHV.pdf	Mozambique	Health
58	2017	Improving electoral performance through citizen engagement in South Africa	https://pdf.usaid.gov/pdf_docs/pa00mx1f.pdf	South Africa	Democracy & Governance
59	2017	Rebuilding interethnic trust in Bosnia and Herzegovina: impact and performance evaluation of USAID/BiH trust, understanding, and responsibility for the future activity (PRO-Future)	https://pdf.usaid.gov/pdf_docs/PA00TFNF.pdf	Bosnia and Herzegovina	Democracy & Governance
60	2017	Business Registration Impact Evaluation - Malawi (IFC/Impact Program)	Non-DEC link	Malawi	Economic Growth
61	2017	Impact evaluation of the feed the future Cambodia helping address rural vulnerabilities and ecosystem stability (HARVEST) project	https://pdf.usaid.gov/pdf_docs/pa00mxgs.pdf	Cambodia	Agriculture & Land
62	2017	Mozambique cell phone savings pilot project endline report	https://pdf.usaid.gov/pdf_docs/pa00mswg.pdf	Mozambique	Agriculture & Land
63	2017	Food and enterprise development (FED) project impact survey	https://pdf.usaid.gov/pdf_docs/PA00MK4N.pdf	Liberia	Agriculture & Land
64	2017	Impact evaluation of USAID Haiti PROJUSTICE program pretrial detention component	https://pdf.usaid.gov/pdf_docs/pa00mz6b.pdf	Haiti	Democracy & Governance
65	2017	Entrepreneurship Status IE - Benin (IFC/Impact Program)	Non-DEC link	Benin	Economic Growth
66	2017	Education for a just society : impact evaluation final report	https://pdf.usaid.gov/pdf_docs/pa00mfpz.pdf	Bosnia and Herzegovina	Democracy & Governance

#	FY	Report Title	DEC URL	Country	Technical Area
67	2017	Evaluation : impact evaluation of the municipal climate change strategies pilot in Macedonia	https://pdf.usaid.gov/pdf_docs/pa00tpqf.pdf	North Macedonia	Environment
68	2017	USAID quality reading project Kyrgyz Republic : final egra and impact report, 2013-2017	https://pdf.usaid.gov/pdf_docs/PA00N67T.pdf	Kyrgyzstan	Education
69	2017	Impact evaluation of the introduction of electronic tax : filing in Tajikistan : endline report	https://pdf.usaid.gov/pdf_docs/pa00tjvv.pdf	Tajikistan	Democracy & Governance
70	2017	Performance & impact evaluation (P&E) of the USAID/Uganda school health and reading program: result 1 school level interventions : impact evaluation final report	https://pdf.usaid.gov/pdf_docs/pbaah456.pdf	Uganda	Education
71	2017	Impact evaluation of the marketing innovation for health project in Bangladesh	https://pdf.usaid.gov/pdf_docs/pa00stvd.pdf	Bangladesh	Health
72	2017	Impact evaluation of the "increasing services for survivors of sexual assault in south Africa" program : endline impact evaluation report	https://pdf.usaid.gov/pdf_docs/pa00mhp2.pdf	South Africa	Democracy & Governance
73	2017	Is LTD's professional development program contributing to improved student learning as measured by standardized tests? : report of an impact evaluation study conducted by the leadership and teacher development (LTD) program in cooperation with the Assessment and Evaluation Department, MoEHE	https://pdf.usaid.gov/pdf_docs/pa00n4jc.pdf	West Bank	Education
74	2017	Impact evaluation of the Rwanda green leaf tea pricing reform : final report	https://pdf.usaid.gov/pdf_docs/pa00w7qn.pdf	Rwanda	Agriculture & Land
75	2017	Midterm impact evaluation of the consolidation and enhanced livelihood initiative : general report [Volume I]	https://pdf.usaid.gov/pdf_docs/pa00mzpd.pdf	Colombia	Economic Growth
76	2017	Final performance and impact evaluation of community mediation through community mobilization project implemented in six Terai districts of Nepal	https://pdf.usaid.gov/pdf_docs/pa00mj3c.pdf	Nepal	Democracy & Governance
77	2017	Impact Evaluation of Malaria Control Interventions on Morbidity and All-Cause Child Mortality in Rwanda, 2000-2010	https://pdf.usaid.gov/pdf_docs/pbaaj732.pdf	Rwanda	Health
78	2017	"Love me, parents!": impact evaluation of a national social and behavioral change communication campaign on maternal health outcomes in Tanzania	Non-DEC link	Tanzania	Health

#	FY	Report Title	DEC URL	Country	Technical Area
79	2017	Malaria Control Interventions Contributed to Declines in Malaria Parasitemia, Severe Anemia, and All-Cause Mortality in Children Less Than 5 Years of Age in Malawi, 2000-2010	https://pdf.usaid.gov/pdf_docs/PBAAJ731.pdf	Malawi	Health
80	2018	Tenure and global climate change (TGCC) : evaluation report	https://pdf.usaid.gov/pdf_docs/PA00T791.pdf	Zambia	Agriculture & Land
81	2018	Final report : hospital accreditation process impact evaluation	https://pdf.usaid.gov/pdf_docs/pa00t5zf.pdf	Indonesia	Health
82	2018	Impact evaluation of the early grade reading activity (EGRA) : final report	https://pdf.usaid.gov/pdf_docs/pa00t3q6.pdf	Malawi	Education
83	2018	Peace through development II : Burkina Faso, Chad, and Niger : impact evaluation endline report	https://pdf.usaid.gov/pdf_docs/pa00swpk.pdf	Burkina Faso, Chad, Niger	Democracy & Governance
84	2018	USAID impact evaluation of the makhalidwe athu project (Zambia) : endline report (final)	https://pdf.usaid.gov/pdf_docs/pa00szjs.PDF	Zambia	Education
85	2018	Impact evaluation of the Niger participatory & responsive governance project : final report	https://pdf.usaid.gov/pdf_docs/pa00w6rs.pdf	Niger	Democracy & Governance
86	2018	Feed the Future Nigeria Livelihoods Project	Non-DEC link	Nigeria	Economic Growth
87	2018	Impact of water users associations on water and land productivity, equity and food security in Tajikistan : final report volume I	https://pdf.usaid.gov/pdf_docs/PA00TDDX.pdf	Tajikistan	Agriculture & Land
88	2018	A ganar alliance impact evaluation : endline report : Honduras	https://pdf.usaid.gov/pdf_docs/pa00t78r.pdf	Honduras	Economic Growth
89	2018	McGovern-Dole International food for education and child nutrition program beoog biiga II endline performance and impact evaluation report	https://pdf.usaid.gov/pdf_docs/pa00tjm3.pdf	Burkina Faso	Health
90	2018	A ganar alliance impact evaluation : endline report : Guatemala	https://pdf.usaid.gov/pdf_docs/pa00t78s.pdf	Guatemala	Economic Growth
91	2018	Basa Pilipinas impact evaluation : final report	https://pdf.usaid.gov/pdf_docs/pa00d5qv.pdf	Philippines	Education
92	2018	Strengthening Tuberculosis Control in Ukraine: Evaluation of the Impact of the TB-HIV Integration Strategy on Treatment Outcomes	https://pdf.usaid.gov/pdf_docs/PA00T89N.pdf	Ukraine	Health
93	2018	Evaluation of Amazonia lee reading intervention in Peru : final report	https://pdf.usaid.gov/pdf_docs/PA00TCQ1.pdf	Peru	Education
94	2018	Evaluation : final report : impact evaluation of the Western Cape emergent literacy intervention in South Africa	https://pdf.usaid.gov/pdf_docs/pa00t61d.pdf	South Africa	Education

#	FY	Report Title	DEC URL	Country	Technical Area
95	2018	Endline impact evaluation : Ghana strengthening accountability mechanisms (GSAM)	https://pdf.usaid.gov/pdf_docs/pa00t6qs.pdf	Ghana	Democracy & Governance
96	2018	Yaajeende final impact evaluation report : an impact evaluation of the yaajeende nutrition-led agriculture program in Senegal (2011-2017)	https://pdf.usaid.gov/pdf_docs/pa00tz2d.pdf	Senegal	Health
97	2018	Impact evaluation : gender and groundnut value chains in Eastern Province, Zambia	https://pdf.usaid.gov/pdf_docs/pa00tcnm.pdf	Zambia	Health
98	2018	The effect of corruption on political behavior in the Peruvian Amazon : impact evaluation of informational campaigns to increase awareness of corruption in politics	https://pdf.usaid.gov/pdf_docs/pa00trjp.pdf	Peru	Democracy & Governance
99	2018	Soapy water handwashing stations : final report	https://pdf.usaid.gov/pdf_docs/pa00srv2.pdf	Kenya	Health
100	2018	Impact evaluation of the women's leadership in small and medium enterprises activity in the Kyrgyz Republic	https://pdf.usaid.gov/pdf_docs/pa00t36f.pdf	Kyrgyzstan	Economic Growth
101	2018	Impact evaluation report : benchmarking a child nutrition program against cash : experimental evidence from an evaluation in Rwanda	https://pdf.usaid.gov/pdf_docs/pa00t9t3.pdf	Rwanda	Health
102	2018	Final report : evaluation of the on-farm water management program : a geospatial impact evaluation of the effects of OFWMP canal improvements on agricultural productivity	https://pdf.usaid.gov/pdf_docs/pa00thdz.pdf	Afghanistan	Agriculture & Land
103	2018	Zimbabwe : works impact evaluation report	https://pdf.usaid.gov/pdf_docs/pa00sxbh.pdf	Zimbabwe	Economic Growth
104	2018	Disseminating Innovative Resources and Technologies to Smallholders, Innovations for Poverty Action: Milestone Report 9	https://pdf.usaid.gov/pdf_docs/PA00N813.pdf	Ghana	Agriculture & Land
105	2018	Can text messages improve local governance? : an impact evaluation of the U-bridge program in Uganda	https://pdf.usaid.gov/pdf_docs/pa00n1tc.pdf	Uganda	Democracy & Governance
106	2018	Final report : evaluation of the infrastructure needs program II : a geospatial impact evaluation of INP II road improvements on economic development	https://pdf.usaid.gov/pdf_docs/pa00tc4z.pdf	West Bank	Economic Growth
107	2018	Final report : impact evaluation of USAID/Nepal's policy dialogue activity (niti sambad) electoral debates and discussions	https://pdf.usaid.gov/pdf_docs/pa00w4sm.pdf	Nepal	Democracy & Governance

#	FY	Report Title	DEC URL	Country	Technical Area
108	2018	Evaluation report : impact and performance evaluation of the USAID Macedonia small business expansion project	https://pdf.usaid.gov/pdf_docs/pa00tffm.pdf	North Macedonia	Economic Growth
109	2018	Multiplex serology for impact evaluation of bed net distribution on burden of lymphatic filariasis and four species of human malaria in northern Mozambique	Non-DEC link	Mozambique	Health
110	2018	Evaluation of the Impact of Malaria Control Interventions on All Cause Mortality in Children Under-five in Angola	Non-DEC Link	Angola	Health
111	2018	Evaluation of the impact of malaria control interventions on all-cause mortality in children under five years of age in Liberia, 2005–2013	https://pdf.usaid.gov/pdf_docs/PA00TFNW.pdf	Liberia	Health
112	2019	DRG learning, evaluation, and research (DRG-LER) activity : impact evaluation of USAID/Georgia's momavlis taoba (MT) civic education initiative (CEI)	https://pdf.usaid.gov/pdf_docs/pa00tq97.pdf	Georgia	Democracy & Governance
113	2019	Final impact evaluation report of the Mandela Washington fellowship programme young African leaders initiative	https://pdf.usaid.gov/pdf_docs/pa00txfx.pdf	Sub-Saharan Africa	Economic Growth
114	2019	Final evaluation of USAID/Cambodia's political processes and reforms program : final report	https://pdf.usaid.gov/pdf_docs/PA00TSQD.pdf	Cambodia	Democracy & Governance
115	2019	Malawi CDCS integrated development impact evaluation : endline report	https://pdf.usaid.gov/pdf_docs/pa00tnp3.pdf	Malawi	Democracy & Governance
116	2019	Training Mentors? Experimental Evidence from Female-Owned Microenterprises in Ethiopia	Non-DEC link	Ethiopia	Economic Growth
117	2019	MUREKE DUSOME IMPACT EVALUATION ENDLINE REPORT Program - Program Impact on Literacy Knowledge, Attitudes, and Practices (KAP)	https://pdf.usaid.gov/pdf_docs/PA00W61H.pdf	Rwanda	Education
118	2019	Impact evaluation : Bangladesh AVC : endline report on impacts associated with the Bangladesh agricultural value chains project	https://pdf.usaid.gov/pdf_docs/pa00trhm.pdf	Bangladesh	Agriculture & Land
119	2019	Rwanda's improved services for vulnerable populations project : impact evaluation : end line report	https://pdf.usaid.gov/pdf_docs/pa00wc3s.pdf	Rwanda	Health
120	2019	Impact evaluation of the women's leadership in small and medium enterprises activity in India	https://pdf.usaid.gov/pdf_docs/pa00tkm1.pdf	India	Economic Growth

#	FY	Report Title	DEC URL	Country	Technical Area
121	2019	Impacts of bt brinjal (eggplant) technology in Bangladesh	https://pdf.usaid.gov/pdf_docs/PA00TZ7Z.pdf	Bangladesh	Agriculture & Land
122	2019	Mindanao youth for development (MYDev) program : FY17 impact evaluation report & FY18/19 (extension) performance evaluation report : measuring youth's employment, perceptions and engagements, and skills	https://pdf.usaid.gov/pdf_docs/pa00w5q2.pdf	Philippines	Education
123	2019	Ethiopia Pastoralist Areas Resilience Improvement and Market Expansion (PRIME) Project Impact Evaluation Endline Survey Report	https://pdf.usaid.gov/pdf_docs/PA00WCWT.pdf	Ethiopia	Environment
124	2019	Data-driven instruction in Honduras : an impact evaluation of the educaccion-pri promising reading intervention : final report	https://pdf.usaid.gov/pdf_docs/pa00wdw1.pdf	Honduras	Education
125	2019	Spring Impact Evaluation Endline Report: Fightback Girls	https://pdf.usaid.gov/pdf_docs/PA00TWMX.pdf	Nepal	Democracy & Governance
126	2019	Impact evaluation of the mayer hashi II project in Bangladesh	https://pdf.usaid.gov/pdf_docs/pa00txvr.pdf	Bangladesh	Health
127	2019	'Voices for peace' impact evaluation of a radio drama to counteract violent extremism in the sahel region in Burkina Faso : endline report	https://pdf.usaid.gov/pdf_docs/pa00w4g3.pdf	Burkina Faso	Democracy & Governance
128	2019	PAHAL RESILIENCE IMPACT EVALUATION: FINAL REPORT	https://pdf.usaid.gov/pdf_docs/PA00WF86.pdf	Nepal	Democracy & Governance
129	2019	Impact evaluation of policies to control deforestation in the Brazilian Amazon : the Paraa state green municipalities program and the federal priority list	https://pdf.usaid.gov/pdf_docs/pa00tkt1.pdf	Brazil	Environment
130	2019	Securing water for food : World Hope impact evaluation : low-cost greenhouse farming in Mozambique	https://pdf.usaid.gov/pdf_docs/pa00tn2n.pdf	Mozambique	Agriculture & Land
131	2019	Impact evaluation of malaria control interventions on morbidity and all-cause child mortality in Mali, 2000-2012	Non-DEC link	Mali	Health
132	2019	SPRING Impact Evaluation Endline Report : Shekina	https://pdf.usaid.gov/pdf_docs/PA00TWMZ.pdf	Ghana	Economic Growth
133	2019	SPRING Impact Evaluation Endline Report: Totohealth	https://pdf.usaid.gov/pdf_docs/PA00TWN1.pdf	Kenya	Economic Growth

ANNEX C: REVIEW INSTRUMENT

Q #	Coding Factor or Question	Coding Instruction
Basic Information about the Impact Evaluation		
I	**DEC ID**	From control sheet
II	**Evaluation Report Title**	From control sheet; review and confirm
III	**Year / Month Published**	From control sheet; review and confirm
IV	**Organizational sponsor of the evaluation**	USAID operating unit: Bureau and Office (e.g., PPL/LER, E3/Land) or Mission (e.g., USAID/Afghanistan)
V	**Primary subject/sector**	Education, health, land tenure, etc.
VI	**Country**	Country name
VII	**Organization that undertook the evaluation**	Name of firm or names of authors not associated with a firm
VIII	**USAID-funded Program** Does the evaluation examine a USAID-funded program?	- Yes - No
IX	**Programming level** About what programming level are the evaluation questions asked?	- A strategy (for a sector, country, etc.: e.g., USAID Water Strategy, USAID/Malawi CDCS) - A project (under which USAID allocates funds to activities; a project-level evaluation might examine outcomes at the project purpose level on a "whole-of-project" basis for a subset of activities) - Activity (one of several activities financed under a project) - Intervention (a specific approach or treatment under an activity, which may involve only one component of that activity) - Pilot (a small activity or component of a larger activity that is explicitly defined as a "pilot" that will test some treatment or intervention to determine whether it works an should be continued or scaled up) - Other (copy and paste description) - Information not provided
X	**Evaluation start date**	Start date (mm, yyyy)
XI	**Evaluation end date**	End date (mm, yyyy)
XII	**Evaluation cost**	Rarely provided but enter if found

Q #	Coding Factor or Question	Coding Instruction
Initial Screening: Evaluation Design Features		
1	**Evaluation design report** Was the evaluation design described in the evaluation report or its annexes?	- The evaluation report included a copy of the approved design in the annex - The evaluation report provided a means to access the design report but did not include it - The evaluation design report description was general, not detailed in nature, and no other version was included or referenced - The evaluation design report was not mentioned
2	**Type of impact evaluation** How did the report describe the type of impact evaluation that was conducted?	- Experimental design (only) *(i.e., randomized)* - Quasi-experimental design (only) *(i.e., researcher manipulation on group assignment but not randomized)* - Quasi-experimental with an experimental design component, *(e.g., encouragement aspect is experimental)* - Non-experimental design *(i.e., no researcher manipulation on group assignment)* - Information not provided
3	**Primary identification strategy** How was assignment to the treatment and comparison groups (corresponding to answering the evaluation questions) established?	- Randomized assignment (i.e., control group) - Cutoff-based assignment (e.g., regression discontinuity design) - Statistical matching (e.g., propensity score matching, other matching approach) - Hand matching against a limited number of characteristics - Convenience assignment (e.g., nearby, easy to reach) - Other (copy and paste description) - Information not provided
4	**Other ID strategies:** Does the evaluation consist of additional identification strategies to answer one or more of the evaluation questions?	- Yes - No
5	*(if Q3 = randomized assignment)* **Randomization technique** Which randomization technique was used?	- Simple randomization using coin toss, lottery, random numbers table, computer-generated random numbers **5a Was the simple randomization done publicly?** - Yes - No - Stratified randomization - Block randomization - Matched pair (units paired then randomized to T/C) - Other (copy and paste description) - Information not provided
6	*(if Q3 = randomized assignment)* **Randomization level**	- Yes, randomization was done at the same level as the intervention itself

Q #	Coding Factor or Question	Coding Instruction
	Was randomization assignment to treatment done at the same unit level as the implementation of the intervention?	- No, randomization was not done at the same level as the intervention itself
7	**Statistical justification for control/comparison group** Does the report provide statistical justification that the control/comparison group is in fact a good comparison for the treatment group prior to the start of the intervention? *[This screening criteria does not apply to experimental designs. For experimental designs, always continue review]*	- Yes, includes a table statistically showing balanced characteristics at baseline **and more than** 10 variables check for balance
		- Yes, includes a table statistically showing characteristics at baseline **but less than** 10 variables check for balance
		- Includes other analysis or detailed justification (text or graphic) claiming the selected control/comparison group is a good comparison, but no statistical table
		- Mentions that analysis to confirm baseline equivalence were conducted but does not provide the information in the report ➔ **STOP REVIEW**
		- No, does not provide any justification that the selected control/comparison group is a good comparison ➔ **STOP REVIEW**
Conceptual Framing		
8	**Evaluation purpose** What is the purpose of the evaluation (i.e., how will findings be used)? *(MARK ALL THAT APPLY)*	- Inform future programming by providing evidence on whether an approach or intervention works - Improve implementation mechanisms of a program by providing evidence of alternative approaches for achieving a given result - Inform the scale-up, replication or continuation of a program/intervention by providing evidence about cost-effectiveness - Other (copy and paste description) - Information not provided
9	**Literature review** Does the report include a relevant literature review of how this evaluation is situated within the existing evidence and in the relevant policy area?	- Yes, provides a relevant literature review clearly describing the existing evidence - Partial, provides a literature review but not described in detail - No, literature review not included
10	**Local context** Does the report explain the local context so that the relevance and external validity of the evaluation can be assessed?	- Yes, provides a description of the local context, including policy and USAID programming - Partial, provides some information of the local context, but not described in detail or about all elements of the intervention(s) being evaluated - No, description of the local context not included
11	**Impact evaluation questions reported** Are the evaluation questions included in the body of the report?	- Yes - No

Q #	Coding Factor or Question	Coding Instruction
11a	Does at least one evaluation question(s) ask about the impact (effect) of a specific intervention on one or more specific outcomes (dependent variables)	- Yes, at least one does (What is the effect of X on Y?) - No
11b	Does at least one evaluation question(s) ask about the relative effectiveness of multiple (alternative) interventions on one or more specific outcome (dependent variables)	- Yes, at least one does (What is the relative effectiveness of A, B and C on Y? or A on Y versus A+B on Y?) - No
11c	How many evaluation questions were asked?	Record the number of questions listed.
11d	Capture the list of evaluation questions included in the report	Copy and paste from evaluation report (if no questions are found but hypotheses to be tested are defined, copy those).
11e	Do the evaluation questions provided in the report match those in the evaluation SOW?	- Evaluation SOW was attached and the evaluation questions/hypotheses in the report and SOW matched - Evaluation SOW was attached but the questions/hypotheses in the report and SOW did not match; however, the report provided an explanation of why they did not match - Evaluation SOW was attached but the questions/hypotheses in the report and SOW did not match, and the report did not provide an explanation of why they did not match - Evaluation SOW was not attached to the evaluation report
12	**Hypotheses** Does the report include research/evaluation hypotheses to be tested?	- Yes - No
12a	How many research/evaluation hypotheses were provided?	Record the number of hypotheses to be tested.
13	**Theory of change** Does the report include a theory of change specific to the evaluation? *[a USAID Mission's results framework or project logical framework may not serve this purpose, as they are sometimes focused on results beyond those the intervention is expected to affect.])*	- Yes, the report included a theory of change diagram or narrative that described the cause-and-effect logic through which the intervention to be examined is supposed to achieve the intended results. - Partial, the report included a theory of change, but it was not a diagram or narrative that described the specific intervention(s) and outcomes on which the evaluation focused. - No, the theory of change diagram or narrative description of the logic model for the intervention(s) was not provided.

Q #	Coding Factor or Question	Coding Instruction
Treatment/Intervention Characteristics		
14	**Treatment operationally defined** How fully was the treatment explained (e.g., frequency of delivery, amount provided)?	- Yes, explanation of the treatment was very detailed and would support an effort to replicate the treatment (e.g., what was delivered, by whom [type of provider/qualifications, e.g., nurse], to whom, in what amount, with what frequency, under what criteria) - Partial, explanation of the treatment was partial, some but not all elements mentioned were described in detail - No, explanation of the treatment was vague and would not be sufficient to support an effort to replicate the treatment
15	**Treatment structure and timing** How was the delivery of the intervention/treatment described?	- Single treatment arm, where units assigned to the treatment group receive the same treatment during a given time period based on design protocol - Multiple treatment arms, where units receive the respective treatment from the arm they are assigned to during a given period based on design protocol. - Phased-in, some units do not receive the treatment in a first period (serve as controls), but receive the treatment in a later period (switch into treatment group) based on design protocol. - Sequencing switched, one group receives Treatment A in the first period and Treatment B in the second; the other group receives Treatment B first and then Treatment A. - Other (copy and paste description) - Information not provided
16	**Treatment complexity** Does this evaluation focus on a single intervention or several bundled activities under the same treatment group?	- Single intervention/component delivered in treatment group (or per treatment arm) - Diverse interventions (e.g., bundled components) within a single treatment group (not separated into treatment arms)
17	**Treatment uniformity** Does the intervention remain consistent and uniform across the treatment group throughout the duration of the evaluation?	- Uniform and consistent intervention across the treatment group - Intervention had minor adjustments or variances within treatment sites - Intervention had major adjustments or variances within treatment sites
Outcome Definitions and Measurement		
18	**Outcomes operationally defined** Did the evaluation provide specific/detailed definitions of each outcome measure the intervention/treatment was expected to affect (e.g., what	- Yes, each outcome the evaluation needed to measure was clearly defined and metrics were provided in design/methods sections in the report or accessible annexes - Yes, each outcome the evaluation needed to measure was clearly defined but metrics were reported in findings section only

Q #	Coding Factor or Question	Coding Instruction	
	terms mean, how the status of outcome is to be measured [i.e., indicators])?	- Partial, outcome measures were vaguely described and/or metrics were not provided - No, outcome definitions and metrics were not provided in the report's design/methods sections.	
19	**Measurement frequency** What is the number of pre-test (baseline) and post-test (follow-up) measurements? *(include midline if it measures outcomes)*	- Pre-tests: _____ Record number of pre-test. If none, record 0.	- Post-tests: _____ Record number of post-test.
20	**Post treatment period** Record measurement timing (in months) after intervention was rolled out.	- Record time (in months) between rollout of intervention(s) and post-test(s). If there are multiple post-tests, record the time period for each one.	
	Sample Size Considerations		
21	**Power calculations** Does the report mention that power calculations were conducted to estimate the needed sample size or MDES?	- Yes, power calculations mentioned and included - Partial, power calculations mentioned but not included - No, not discussed	
22	**Sample size** Is the sample size needed to conduct the IE reported?	- Yes, sample size reported for treatment and comparison/control groups, either or both - Partial, sample size only vaguely described, e.g., large - No, sample size not described	
22a	Record the required sample size for treatment and comparison/control groups	Record the required sample size as reported (overall sample size or sample size for each treatment and comparison/control groups)	
23	**Power** Is the power level on which sample size was based reported?	- Yes, power calculation and level on which sample was based is provided - Partial, power discussed only vaguely (e.g., adequate power achieved) - No, power not discussed	
23a	Record power level reported if provided	- .80 - .90 - Other (if other, record power level on which sample size was based)	
24	**Minimum detectable effect size (MDES)** Is the MDES used in power calculation provided (*ex-ante* effect size)?	- Yes, MDES in power calculation (based on Cohen's d) provided (e.g., 0.2, 0.5, 0.8) - Partial, MDES only vaguely described (e.g., small, medium, large) - No, MDES not discussed	
24a	Capture MDES reportedly used in the power calculation if provided	Record the MDES used (e.g., 0.2, 0.5)	
24b	Did the report provide the basis for the MDES used in the power calculation?	- Yes, report explained the basis for MDES it used (e.g., policy objective, project target, results from studies on similar programs, ex ante simulation, etc.)	

Q #	Coding Factor or Question	Coding Instruction
		- No, report did not discuss the basis for MDES it used
25	**Expected take-up** Did the report explain and account for the expected treatment take-up in the power calculations?	- Yes, report explained the expected take-up mathematically (e.g., 65% take-up) and adjusted sample size to account for it - Partial, report explained the expected take-up and/or did not adjust the sample size to account for it - No, report did not discuss the expected take-up
26	**Expected attrition** Did the report explain and account for the expected attrition in the sample size?	- Yes, report explained expected attrition mathematically (e.g., 10% drop out) and adjusted sample size to account for it - Partial, report vaguely explained expected attrition (e.g., large, modest) and how it affected sample size calculations (e.g., attrition expectations were factored in) - No, report did not discuss expected attrition - Attrition is not a relevant factor for this evaluation (e.g., cross-sectional cluster design)
27	**Is the evaluation a cluster-level or individual-level design?**	- Cluster level - Individual level
28	*(If the evaluation is a cluster design)* **Cluster specification** Does the report specify and describe the cluster unit of assignment (e.g., schools, or districts) and the unit of analysis (e.g., students)?	- Yes, clusters were described/explained in detail - No, clusters were not described/explained in detail - Not a cluster design
29	*(If the evaluation is a cluster design)* **Intra-cluster correlation** Does the sample size calculation account for intra-cluster correlation (ICC)?	- Yes, ICC included in power calculation - Partial, ICC only vaguely described (e.g., small, medium, large) - No, ICC not discussed - Not a cluster design
29a	Did the report provide the basis for the ICC used in the power calculation?	- Report explained the basis for ICC it used (e.g., empirical data) - Basis for ICC used in power calculation not discussed - Not a cluster design
Data Collection and Analysis		
30	**Data collection description** Does the report describe the data collection methods used to answer the main IE question(s)?	- Yes, data collection methods used were fully described in the report or its annexes - Partial, data collection methods were discussed but not in detail either in the report or in an annex - No, data collection methods not discussed
30a	**Data collection methods used**	- Surveys - Administrative data

Q #	Coding Factor or Question	Coding Instruction
	What were the data collection methods used in the evaluation? *(MARK ALL THAT APPLY)*	- Individual interviews - Group interviews - Focus groups - If other, cut and paste in descriptions of other methods
30b	Are data collection instruments provided as evaluation report annexes?	- Yes, data collection instruments included in the annex - No, data collection instruments not included
31	**Pre-analysis plan** Was the pre-analysis plan included in the evaluation report or its annexes?	- Yes, the evaluation report included the pre-analysis plan in the annex - Yes, the evaluation report referenced and described the pre-analysis plan and provided a means to access it but did not include it - Partial, the evaluation report includes a general and vague description of the pre-analysis plan and does not include it or reference it - No, pre-analysis plan not discussed
32	**Data analysis methods description** Does the report describe the analysis undertaken to estimate impact effects?	- Yes, report provided a full and detailed description of the actual analysis undertaken, including model specifications and tests run to determine impact estimates, group differences, etc. - Partial, actual analysis discussed but not displayed in detail in report - No, actual analyses conducted were not discussed (e.g., findings were presented but not how they were arrived at)
32a	**Analysis methods used** What were the estimation methods used in the evaluation? *(Mark all that apply)*	- Single difference (comparison of means) - Difference-in-differences - Regression estimation - Sensitivity analyses – including reporting of multiple model specifications and related results and effect of pooled and non-pooled standard errors, where appropriate - If other, cut-and-paste the descriptions of analysis methods used including specific tests, etc.
33	**Heterogeneity of impacts** Does the evaluation conduct any sub-group analysis to assess how the impact estimates differ by groups?	- Yes, the evaluation conducts and explains the sub-group analysis (e.g., by beneficiary group, by dosage) - No, the evaluation does not conduct any sub-group analysis but explains why it was not done - No, the evaluation does not conduct or mention any sub-group analysis
33a	*(If Q33=Yes)* **Gender**	- Yes - No

Q #	Coding Factor or Question	Coding Instruction
	Does the evaluation assess how impact estimates differ by gender?	
34	**Ethics** Does the report mention that ethics approval from an authorized institution and informed consent from respondents were obtained before the start of data collection?	- Both, ethics approval and informed consent obtained - Only ethics approval obtained - Only informed consent obtained - Neither ethics approval nor informed consent obtained
	Common Threats to Validity	
35	**Treatment fidelity** Does the report describe the extent to which the intervention(s)/treatment(s) was/were delivered as planned?	- Report stated that the treatment was delivered as planned, including verifying treatment fidelity with respect to the amount provided, the frequency, who provided it, and other aspects of the treatment's operational definition - Report stated that the treatment was provided as planned but did not discuss in detail its fidelity to the design protocol - Report states that treatment was not provided as planned - Intervention/treatment's actual delivery not described or discussed
36	**Actual treatment take-up** Does the report describe the actual take-up of the intervention(s)/treatment(s) by those who were intended to receive the intervention?	- Yes, the report stated the actual treatment take-up (e.g., 65% take-up) and described the reasons for the imperfect compliance - Partial, the report mentions treatment take-up in a vague manner without a clear discussion of it - No, the report did not discuss the actual take-up of the intervention in the treatment group
37	**Contamination** Did the evaluation report that people or sites that were intended to be in the control/comparison group received the intervention?	- The report indicates that no cases of control/comparison group units received the intervention or part of it - The report indicated that individuals/sites in the comparison/control group received the intervention/treatment they were not intended to receive. The extent of this distortion was described in detail and accounted for in the analysis. - The report indicated this deviation occurred but does not state the extent of this deviation or does not account for it in the analysis - The report did not discuss whether this deviation occurred *[This does not mean it did not happen, only that it was not discussed].*
37a	Does the report provide an explanation for how	- Members of the control/comparison group tried [successfully] to access the treatment even if it was not offered to them

Q #	Coding Factor or Question	Coding Instruction
	comparison/control units received the intervention?	- Implementer error or deviation from study protocol made the treatment available to control/comparison group members or sites - The treatment is assigned based on a continuous eligibility index, but the eligibility cutoff is not strictly enforced (applies to RDD design) - Selective migration, where individuals from comparison group sites choose to move to another site receiving the treatment. - Explanation of how reported cases of control/comparison groups received treatment was too vague to determine mechanism involved (copy and paste description) - Explanation of how comparison/control groups accessed the treatment was not discussed
38	**Actual attrition** Does the evaluation report the actual attrition and assess how it affects the evaluation findings?	- Yes, actual attrition is reported in explicit terms (e.g., 10% drop out) and report shows whether attrition was random or systematic (i.e., if it is correlated with the treatment or threats balance of the treatment and control/comparison groups) - Partial, actual attrition reported or discussed but threat to validity of the evaluation findings was not assessed - No, actual attrition was not discussed *[This does not mean it did not happen, only that it was not discussed].*
39	**Missing data/non-response** Does the report discuss missing data/non-response?	- Yes, non-response rate reported in explicit terms (e.g., 5%) and reasons for non-response or missing data described - Partial, non-response mentioned in a vague manner without a precise rate and no discussion of the reasons for it - No, non-response rate not discussed *[This does not mean it did not happen, only that it was not discussed].*
Reporting of Findings		
40	**Reported sample size** Are sample sizes for the treatment and control groups reported in the findings?	- Yes - No
41	**Format of reported impact estimates** How does the evaluation report the impact estimates?	- Impact estimate was reported as a standardized effect size (i.e., in terms of standard deviations) - Impact estimate was reported as an absolute or percentage change in the status of the outcome variables (e.g., percent change in treatment group versus comparison/control group) - Impact estimate was not reported in numerical terms but was described (e.g., large impact) - If other, cut and paste how impact was reported

Q #	Coding Factor or Question	Coding Instruction
42	**Statistical Significance** Does the report include an outputs table (or text) with the impact estimates and their statistical significance using conventional levels?	- Yes, reported the confidence intervals around point estimates and/or significance level (e.g., 0.05) - Partial, reported statistical significance but not the actual confidence intervals or significance level (e.g., only states that a finding was significant) - No, statistical significance of findings not reported
43	**Practical significance** Does the evaluation discuss the practical significance of the impact estimate, in terms of the implications of the size/sign of the effect?	- Yes, practical significance of impact estimate was clearly explained (magnitude of effect size and implications) - Partial, practical significance mentioned but not described (e.g., the study finding suggests that farmers could improve their crop yields by adopting this practice) - No, practical significance not discussed
44	**"No impact" discussion** Does the report include a discussion of why effect was not detected (e.g., under-powered, implementation issues, flawed program design)?	- Yes, the evaluation included this type of discussion - No, it did not unpack the "no impact" result to provide decision-makers with some understanding of "why" no impact was detected - This is not a case of "no impact" result
45	**Study limitations** Did the impact evaluation report include a statement of the study limitations in the body of the report prior to presenting its findings?	- Yes - No
46	**Report Structure** Did the report structure the findings section(s) in relation to evaluation questions, as opposed to presenting information in relation to project objectives or in some other format?	- Yes - No
47	**Conclusions/ Recommendations** Did the report include a conclusions and/or recommendations section advising USAID decision-makers on the implications of the impact evaluation's results for continuing/discontinuing, replicating, or scaling up the intervention(s)?	- Yes, the evaluation included this type of discussion - No, the evaluation did not provide decision-makers with conclusions or recommendations that explain the indications for action based on the impact evaluation's results - N/A, no recommendations included

Q #	Coding Factor or Question	Coding Instruction
47a	**Recs supported by Findings/Conclusions** Are all the recommendations supported by the findings and conclusions presented? *(Can the reader follow a transparent path from findings to conclusions to recommendations?)*	- Yes - No - N/A, no recommendations included
Cost-Effectiveness		
48	**Cost-effectiveness question** Do the evaluation question(s) ask about the cost-effectiveness of the intervention relative to current practice or one or more alternatives?	- Yes, at least one evaluation question does - No
49	**Cost information** Did the report state the cost of delivering the program/ intervention being evaluated?	- Yes, cost data of program/intervention being evaluated was provided - No, but cost data for the broader activity was provided - No, cost data was not provided
49a	Record cost of delivering the program/intervention being evaluated	Record dollar amount if provided
50	**Cost per unit** Does the report state the cost per unit/beneficiary of delivering the program/intervention being evaluated?	- Yes, the cost per unit information was provided - No, cost per unit information was not provided
50a	Record cost per unit/beneficiary	Record dollar amount if provided
51	**Cost-effectiveness results** Is the cost-effectiveness of the program/intervention reported and discussed?	- Yes, the report describes in detail how cost effective the treatment (intervention) is in comparison to current practice or some other relevant alternative (e.g., the government would save XX by adopting this program to providing vaccinations) - Partial, the report says the treatment is cost effective but does not provide evidence to support that claim - No, cost effectiveness of the treatment is not discussed
51a	Record the cost effectiveness amount (e.g., $ per unit change in outcome OR additional outcome gained per $ spent)	Record the cost-effectiveness value reported

ANNEX D: FINDINGS FOR EACH QUALITY ELEMENT