USAID
FROM THE AMERICAN PEOPLE

OFFICE OF EDUCATION

Tool: Assessing the Quality of Education Evaluations

The evaluation quality assessment tool was developed as part of the Assessment of the Quality of USAID-Funded Evaluations in the Education Sector, 2013-2016. For more information about the content of this document, contact Elena Walls at ewalls@dexisonline.com.

Version: December 2017

Overview

Purpose

This quality assessment of USAID-funded evaluations in the education sector was commissioned by USAID's Office of Education (E3/ED) to support a larger effort to curate, analyze, and disseminate robust evidence related to the objectives of USAID's 2011-2015 Education Strategy. The assessment achieves two related objectives:

- Identifies strengths and weaknesses in the quality of USAID-funded education evaluations
- Codifies best practices in designing, implementing and reporting on evaluations for use by USAID Missions, partners, and the education sector at large, as appropriate.

The results of this assessment also inform a related evaluation synthesis study, commissioned by E3/ED, which summarizes findings and lessons learned from Education Strategy-related evaluations published between 2013 and 2016. Only the education evaluations that met quality standards through this assessment were included in the subsequent evaluation synthesis study.

Intended Audience

The findings and the assessment tool are intended for USAID Missions, USAID Washington and USAID partners involved in evaluating USAID-funded education programming. The tool could also be useful for other sectors of international development in helping improve the strength of evidence generated by evaluations.

How to Use the Assessment Tool

The evaluation quality assessment tool developed as part of the quality review process is based on internationally recognized best practices in evaluation, USAID policies, and standards in the evaluation practice in the US. The tool is structured around seven aspects of quality that can be applied to any evaluation, regardless of type. Not all items in the assessment tool will be applicable to all evaluations; they should be used as relevant and appropriate, depending on the evaluation parameters. Stakeholders might want to use the framework and tool differently; for example, it might be used by:

- USAID Missions when procuring and managing evaluations;
- Partners when planning and implementing evaluations;
- The sector at large when learning about evaluation practice and curating the evidence generated through evaluations.

E3/ED is planning on implementing a review of evaluation quality on a periodic basis, as part of the larger evaluation synthesis study.

TABLE 1. Assessment Tool[1]

PRINCIPLE OF QUALITY	IMPACT EVALUATIONS	PERFORMANCE EVALUATIONS		REVIEW RESULT	OVERALL CONCLUSION	NOTES/ JUSTIFICATION
		QUANTITATIVE	QUALITATIVE			
CONCEPTUAL FRAMING	[1] Are the research/evaluation questions included in the report?			yes/no	adequate/not adequate	
	[2] Does the report include research/evaluation hypotheses?			yes/no		
	[3] Are the evaluation questions appropriate for the intervention's conceptual framework (logframe/theory of change/ results framework)?			yes/partial/no/ not applicable		
	[4] Does the report acknowledge/draw upon existing relevant research?			yes/partial/no		
	[5] Does the report explain the local context in sufficient detail?			yes/partial/no		
OPENNESS AND TRANSPARENCY	[6] Is the report open about study limitations with the implementation of the **evaluation**, such as issues faced during data collection that might affect the study's design?			yes/partial/no	adequate/not adequate	
	[7] Is the report open about study limitations due to issues with the implementation of the **intervention** being evaluated?			yes/partial/no/ not applicable		
	[8] Does the discussion about the findings refer to relevant contextual factors or methodological considerations?			yes/no/not applicable		
	[9] Is the report open about potential influence due to the study team composition?			yes/partial/no		
CULTURAL APPROPRIATENESS	[10] Does the report list steps taken to ensure that study questions and methodology are informed by local stakeholders, are culturally relevant and contextually appropriate?			yes/no	adequate/not adequate	
	[11] Does the report list steps to address and document that data collection tools were developed/adapted with participation of relevant local stakeholders and are culturally appropriate?			yes/partial/no		
	[12] Does the report list steps taken to validate findings/conclusions/recommendations with local stakeholders **as part of the evaluation**?			yes/no		

[1] The evaluation quality assessment tool developed as part of the Assessment of the Quality of USAID-Funded Evaluations in the Education Sector, 2013-2016, was revised upon completion of the review, based on the comments from reviewers. This version reflects these revisions. The tool's framework is based on Assessing the Strength of Evidence in the Education Sector (DFID, 2015; produced by Building Evidence in Education Donor Working Group).

PRINCIPLE OF QUALITY	IMPACT EVALUATIONS	PERFORMANCE EVALUATIONS		REVIEW RESULT	OVERALL CONCLUSION	NOTES/ JUSTIFICATION
		QUANTITATIVE	QUALITATIVE			
	[13] Was the study designed to take into account locally relevant stratifiers, such as political, social, ethnic, religious, geographical or sex/gender phenomena during data collection and data analysis?			yes/partial/no		
ROBUSTNESS OF METHODOLOGY	[14] Is the methodology explained in sufficient detail?			yes/partial/no	adequate/not adequate	
	[15] Is the methodology appropriate for answering posed study questions?			yes/partial/no/ not applicable		
	[16] Does the counterfactual meet standards of rigor?			yes/no/not applicable		
	[17] Does the report include information from multiple data sources and how the data were triangulated?			yes/partial/no/ not applicable		
	[18] Does the report mention steps taken to mitigate common threats to the integrity of the evaluation (such as non-equivalence at baseline, non-compliance, spillover, systematic attrition) or common biases (confounding bias, selection bias, experimenter bias, etc)?			yes/partial/no		
	[19] For the **quantitative** research methods used, are the sampling approach and sample size calculations presented in sufficient detail (to include, at a minimum, type of analysis, MDES, alpha and beta)?			yes/partial/no/ not applicable		
			[20] For the **qualitative** research methods used, is the sampling approach described in sufficient detail? (at a minimum, a rationale for the sample size and method of sample selection) and is it appropriate for the study objectives?	yes/partial/no/ not applicable		
VALIDITY	[21] Do indicators used in the evaluation capture the construct or phenomenon being investigated?			yes/partial/no/ not applicable	adequate/not adequate	
	[22] Were the sampling conducted in such a way such that the results are generalizable to the population of beneficiaries reached through the activity?			yes/partial/no/ not applicable		

PRINCIPLE OF QUALITY	IMPACT EVALUATIONS	PERFORMANCE EVALUATIONS		REVIEW RESULT	OVERALL CONCLUSION	NOTES/ JUSTIFICATION
		QUANTITATIVE	QUALITATIVE			
	[23] Does the report allude to whether the study findings may have been biased by the activity of doing the study itself?			yes/no		
	[24] Does the report address the external validity of findings?			yes/partial/no/ not applicable		
	[25] Were all data collection tools piloted with representatives of target populations prior to beginning of the data collection?			yes/partial/no		
	[26] Are confidence intervals reported around point estimates?			yes/no/not applicable		
	[27] Is treatment effect presented in terms of effect size?			yes/no/not applicable		
RELIABILITY	[28] Does the report list steps taken to ensure that data were collected with a high degree of reliability?			yes/partial/no	adequate/not adequate	
	[29] Does the report adequately address missing data/non-response?			yes/partial/no		
COGENCY	[30] Are all the study questions, including sub-questions, answered?			yes/no/not applicable	adequate/not adequate	
	[31] Does the Executive Summary include answers to all of the study questions?			yes/no		
	[32] Is the report accessible to the audiences for whom the report indicates it is written (e.g., minimizing technical jargon if intended to the general public)?			yes/no		
	[33] Are conclusions based on findings and are the findings related to the evaluation questions?			yes/partial/no/ not applicable		
	[34] Is the narrative in the report supported by charts, maps and infographics that help non-technical audiences easily understand the study findings?			yes/partial/no		

TABLE 2. Item Description and Source

PRINCIPLE/ABBREV. ITEM	QUESTION	SCORE	DESCRIPTOR	SOURCE
Conceptual Framing: Study questions included	[1] Are the research/evaluation questions included in the report? **[All evaluation types]**	yes/no	All research/evaluation questions must be phrased as questions; it is not enough that they be inferable from the stated objectives of the study. Questions must be clearly stated and be answerable through the reported research methods.	ADS 201maa: Evaluation reports should adequately address all evaluation questions included in the SOW, or the evaluation questions subsequently revised and documented in consultation and agreement with USAID.
Conceptual Framing: Study hypotheses included	[2] Does the report include research/evaluation hypotheses? **[All evaluation types]**	yes/no	Research/evaluation hypotheses must be explicitly described; it is not enough that they be inferable from the stated objectives of the study.	BE2, Checklist: Does the study outline a hypothesis?
Conceptual Framing: Study questions appropriate given the intervention's conceptual framework	[3] Are the evaluation questions appropriate for the intervention's conceptual framework (logframe/theory of change/ results framework)? **[IE, Perf. Quant, Perf. Qual]**	yes/partial/no /NA	All research/evaluation questions should be based on the intervention's conceptual framework. "Partial" score could be given when some, but not all, listed evaluation questions correspond to the intervention's conceptual framework. "NA" score should be given to research studies that do not evaluate a specific intervention.	BE2, Checklist: Does the study pose an appropriate research question?
Conceptual Framing: Study acknowledges/draws upon existing country-specific research	[4] Does the report acknowledge/draw upon existing relevant research? **[All evaluation types]**	yes/partial/no	Studies should build on existing research, both local and funded by international donors. The report should specify how questions, methodology, tools and analysis plans are informed by prior research. "Partial" score could be given when only some of the questions are informed by existing knowledge.	BE2, Checklist: Does the study acknowledge existing research?

PRINCIPLE/ABBREV. ITEM	QUESTION	SCORE	DESCRIPTOR	SOURCE
Conceptual Framing: Local context provided allows non-experts appreciate relevance of the study	[5] Does the report explain the local context in sufficient detail? **[All evaluation types]**	yes/partial/no	The local context should be explained in enough detail for a general audience to be able to appreciate the relevance of the intervention being evaluated. "Partial" score could be given when some, but not all, elements of the intervention have corresponding contextual information.	USAID Evaluation Policy, page 8: Evaluation reports should include sufficient local and global contextual information so that the external validity and relevance of the evaluation can be assessed.
Conceptual Framing: Conclusion	Conceptual framing: Conclusion **[All evaluation types]**	adequate/not	**Adequate:** Overall, this evaluation demonstrates adherence to principles of conceptual framing **Not Adequate:** This evaluation contains major deficiencies in demonstrating adherence to principles of conceptual framing or provides insufficient information for determining this	
Conceptual Framing: Justification	Conceptual framing: Notes/Justification **[All evaluation types]**		For instance: *"The authors acknowledge existing research and make clear how their analyses sit within the context of existing work. They provide a theoretical framework in the report, where they outline their major assumptions. The study also poses specific research questions."*	
Openness and Transparency: Open about limitations to implementing the study	[6] Is the report open about study limitations with the implementation of the **evaluation**, such as issues faced during data collection that might affect the study's design? **[All evaluation types]**	yes/partial/no	Limitations to the implementation of the evaluation should be clearly presented. Clarity around study limitations is particularly important if they directly impact the evaluator's ability to credibly and effectively answer an evaluation question or impact generalizability of the findings (i.e., if data collection was successful but more expensive or inconvenient than anticipated, it is not a limitation). "Partial" score could be given if the report mentions limitations without discussing them in detail.	BE2, page 17: The study should also clearly state the sample size.
Openness and Transparency: Open about limitations to implementing the intervention	[7] Is the report open about study limitations due to issues with the implementation of the **intervention** being evaluated? **[IE, Perf. Quant, Perf. Qual]**	yes/partial/no /NA	Limitations to the implementation of the intervention being evaluated should be clearly presented, such as delays or changes that may compromise the integrity of the evaluation design. "Partial" score could be given if the report mentions imitations without discussing them in detail. "NA" score should be given to research studies that do not evaluate a specific intervention.	BE2, page 17: An important sign of quality is whether the author is being self-critical; being open about limitations.

PRINCIPLE/ABBREV. ITEM	QUESTION	SCORE	DESCRIPTOR	SOURCE
Openness and Transparency: Alternative interpretations of the findings included	[8] Does the discussion about the findings refer to relevant contextual factors or methodological considerations? **[All evaluation types]**	yes/no/NA	The evaluation report should balance the presentation of the findings with a discussion contextualizing them and/or addressing how they might be affected by methodological decisions. This discussion might include broaching alternative explanations for the findings. If some findings yield inconsistencies with others, this should be discussed as well. "NA" score should be given if individually findings were not conducive with discussion about contextual or methodological considerations and collectively they are not contradictory.	BE2, page 17: An important sign of quality is whether the author is being self-critical; being open about (...) alternative interpretations and pointing out inconsistencies with other results.
Openness and Transparency: Open about potential biases due to the study team composition	[9] Is the report open about potential influence due to the study team composition? **[All evaluation types]**	yes/partial/no	USAID encourages study teams to include at least one evaluation specialist, host country team members, and a team leader who is external to USAID. USAID also requires that evaluation team members certify their independence by signing statements **disclosing any conflict of interest or fiduciary involvement with the project or program** they will evaluate. It is expected that an evaluation will indicate that such forms, or their equivalent, are on file and available or are provided in an evaluation annex. "Partial" score could be given if some, but not all, these recommendations are followed.	BE2, Checklist: Does the researcher acknowledge their own subjectivity in the process of the research?
Openness and Transparency: Conclusion	Openness and transparency: Conclusion **[All evaluation types]**	adequate/not	**Adequate:** Overall, this evaluation demonstrates adherence to principles of openness/transparency **Not Adequate:** This evaluation contains major deficiencies in demonstrating adherence to principles of openness/transparency or provides insufficient information for determining this	
Openness and Transparency: Justification	Openness and transparency: Notes/Justification **[All evaluation types]**		For instance: *"The authors are transparent about the design and methods that have been employed in the evaluation as well as the data (and resulting sample) that have been gathered and analyzed. This allows for the study to be repeated and corroborated."*	

PRINCIPLE/ABBREV. ITEM	QUESTION	SCORE	DESCRIPTOR	SOURCE
Cultural Appropriateness: Study questions and methodology informed by local stakeholders	[10] Does the report list steps taken to ensure that study questions and methodology are informed by local stakeholders, are culturally relevant and contextually appropriate? **[All evaluation types]**	yes/no	The evaluation questions and methodology should be informed by relevant local stakeholders. This could be done during in-country design workshops as well as through meeting with the ministry or other relevant stakeholders.	ADS 201sae: Is there reasonable assurance that the data collection methods being used do not produce systematically biased data.
Cultural Appropriateness: Data collection tools developed with participation of local stakeholders	[11] Does the report list steps to address and document that data collection tools were developed/ adapted with participation of relevant local stakeholders and are culturally appropriate? **[All evaluation types]**	yes/partial/no	The report should describe whether tools have been developed to suit the local context, such as whether the tool was developed by international experts and then merely translated into a local language or whether local knowledge has been used effectively in the adaptation of the tool to reflect resources relevant to the context, such as including support from host country experts. Quality control of translators (back-translation) is recommended. "Partial" score could be given if some, but not all tools suit the local context.	BE2, page 20: For all research designs, it is important to consider the extent to which the measures/ instruments/ variables used in the study suit local contexts. The reviewer should note whether measures have been developed to suit the local context: does the study, for instance, merely translate into a local language or recognize that a test developed in a specific linguistic area may not be automatically suitable to a local context with translation or because of multiple socio-linguistic processes? The reviewer should also note whether local knowledge has been used effectively in the adaptation of measures to reflect resources relevant to the context; for example, are the instruments designed with support and recognition from the local community?

PRINCIPLE/ABBREV. ITEM	QUESTION	SCORE	DESCRIPTOR	SOURCE
Cultural Appropriateness: Findings/conclusions/recommendations validated with local stakeholders	[12] Does the report list steps taken to validate findings/conclusions/ recommendations with local stakeholders **as part of the evaluation**? **[All evaluation types]**	yes/no	Findings, conclusions and recommendations must be communicated to the appropriate audiences in a culturally and contextually suitable way prior to finalization of the report, in order to validate accuracy of conclusions and help inform recommendations. Steps to validate these with local stakeholders may include in-country presentations and workshops conducted during the evaluation (instead of as dissemination studies after the evaluation was concluded).	EGRA Toolkit, 2nd edition, page 122: Results must be communicated to the appropriate audiences in a culturally and contextually suitable way in order to support understanding and action.
Cultural Appropriateness: Findings disaggregated by locally relevant stratifiers	[13] Was the study designed to take into account locally relevant stratifiers, such as political, social, ethnic, religious, geographical or sex/gender phenomena during data collection and data analysis? **[All evaluation types]**	yes/partial/no	The extent to which a study takes into account locally relevant stratifiers has considerable bearing on the study's design, its analytical strategy and the interpretation of its findings. Being informed by locally relevant stratifiers might include making cross-cultural or cross-linguistic comparisons part of the analytical strategy or ensuring that knowledge of the local context is used in the interpretation of differential effects between groups. "Partial" score should be assigned when the study is purposeful with considering variable impacts on gender but not any other stratifiers.	BE2, page 20: This includes the extent to which the analysis includes locally relevant social stratifiers (for example, socio-economic status, gender, rural-urban differences, etc.) and influences which may affect interpretation of results.
Cultural Appropriateness: Conclusion	Cultural appropriateness: Conclusion **[All evaluation types]**	adequate/not	**Adequate:** Overall, this evaluation demonstrates adherence to principles of cultural appropriateness. **Not Adequate:** This evaluation contains major deficiencies in demonstrating adherence to principles of cultural appropriateness or provides insufficient information for determining this.	

PRINCIPLE/ABBREV. ITEM	QUESTION	SCORE	DESCRIPTOR	SOURCE
Cultural Appropriateness: Justification	Cultural appropriateness: Notes/Justification **[All evaluation types]**		For instance: *"The evaluation describes systematic processes used to check for the cultural relevance of measurement items (for example, in the absence of lists of age-specific words for Bangla-speaking children, a list was created of words that fit two criteria: they should be known to grade 1 or 2 children but unknown to preschoolers, and they should be used in the storybooks). Thus, the instrument used is culturally sensitive. The analysis is also culturally sensitive, as it discusses the factors that undermine or promote educational outcomes within the Bangladeshi context. The study discusses the use of two supply-and-demand side interventions – a school-only grant and a school grant plus an education allowance – which the authors discuss in relevance to the context, where grants are used to provide key inputs to schools while the education allowance provides a conditional monetary incentive for out-of-school children to attend school."*	
Robustness of Methodology: Methodology explained in detail	[14] Is the methodology explained in sufficient detail? **[All evaluation types]**	yes/partial/no	USAID requires that an evaluation report identifies the study design, data collection methods and data analysis techniques used. It is common to include the methodology description in the body of the report under a methodology section with a longer and more detailed methods annex. The description of methods must indicate: how respondents were selected, • what types of interviews were conducted; • with whom they were conducted (e.g., key informant interviews, individual interviews with beneficiaries, group interviews) and; • detailed information on the kinds of analyses that were conducted (e.g., correlations, regressions, content analysis, pattern analysis). "Partial" score could be given if some, but not all elements mentioned (design, data collection methods and data analysis techniques) were described in sufficient detail.	ADS 201maa: Evaluation methodology should be explained in detail and sources of information properly identified.

PRINCIPLE/ABBREV. ITEM	QUESTION	SCORE	DESCRIPTOR	SOURCE
Robustness of Methodology: Methodology appropriate for answering posed study questions	[15] Is the methodology appropriate for answering posed study questions? **[All evaluation types]**	yes/partial/no /NA	USAID recognizes that the methodology used to address the posed questions may be defined in the issued Scope of Work for the evaluation. USAID also recognizes that different designs are more or less appropriate to answering different research questions, and that the selection of method (or methods) for a particular evaluation also balances cost, feasibility, and the level of rigor needed to inform specific decisions. Assessing the appropriateness of the chosen methodology may be further complicated when the evaluation includes a variety of questions that require a mixed-method approach; for such evaluations, the assessment of the methodology must include the review of the evaluation design vis-a-vis each stated study questions. "Partial" score could be given if the methodology proposed is appropriate for some, but not all posed questions. "NA" score should be given if the study does not pose research/evaluation questions.	USAID Evaluation Policy, page 8: evaluation should principally consider the appropriateness of the evaluation design for answering the evaluation questions as well as balance cost, feasibility, and the level of rigor needed to inform specific decisions.
Robustness of Methodology: Counterfactual meet standards of rigor	[16] Does the counterfactual meet standards of rigor? **[IE]**	yes/no/NA	Measuring what would have happened in the absence of an intervention is a requirement for establishing a causal relationship. A counterfactual can be created in a number of ways, from simply using respondents from a geographically close unit as comparison group to using statistical analysis to compensate for the potential selection biases of non-randomization to randomly assigning subjects to treatment(s) and control groups. Considerations about its rigor may include a review of information in the report about baseline equivalence, differential attrition, etc. "NA" score should be given if the evaluation is not an Impact Evaluation.	USAID Evaluation Policy, page 3: 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 the observed change. (...) Performance evaluations encompass a broad range of evaluation methods. They often incorporate before-after comparisons, but generally lack a rigorously defined counterfactual.

PRINCIPLE/ABBREV. ITEM	QUESTION	SCORE	DESCRIPTOR	SOURCE
Robustness of Methodology: Data triangulation described as part of methodology	[17] Does the report include information from multiple data sources and how the data were triangulated? **[All evaluation types]**	yes/partial/no /NA	Typically, stronger bodies of evidence are likely to emerge if similar findings are obtained from different types of data (e.g., tests, interviews, observations) and respondent types (e.g., students, parents, teachers). It is important that contradictory data be taken into account when discussing the findings. "Partial" score could be given if data from different sources are presented but the findings don't connect them into a coherent narrative. "NA" score should be given if the evaluation does not use multiple data sources.	CASP, Qualitative Checklist: To what extent contradictory data are taken into account?
Robustness of Methodology: Addressed internal validity, either threats to inference or common biases	[18] Does the report mention steps taken to mitigate common threats to the integrity of the evaluation (such as non-equivalence at baseline, non-compliance, spillover, systematic attrition) or common biases (confounding bias, selection bias, experimenter bias, etc)? **[All evaluation types]**	yes/partial/no	USAID Evaluation Policy requires that evaluation reports address methodologically common limitations, such as when there is a disjunction between the treatment that is assigned and the treatment that is received (non-compliance). "Partial" score could be given if some, but not all threats or biases identified are discussed.	USAID Evaluation Policy, page 10: Evaluation reports that include the original statement of work, a full description of methodology (or methodologies) used, as well as the limitations in the inferences that can be drawn.
Robustness of Methodology: Described sampling approach and parameters used to compute sample size	[19] For the **quantitative** research methods used, are the sampling approach and sample size calculations presented in sufficient detail (to include, at a minimum, type of analysis, MDES, alpha and beta)? **[IE, Perf. Quantitative]**	yes/partial/no /NA	Details of power calculation should be included in either the main body of the report or in an annex. This should include the parameters used in the power function that relates power (beta) to its determinants: (1) level of significance (alpha), (2) minimum detectable effect size (MDES) or minimum detectable impact (MDI), (3) and the sample size. "Partial" score could be given if the description of the sample size calculations presents only some of the parameters used. "NA" score could be given if the evaluation/research used only qualitative research methods	JPAL's Running Randomized Evaluations, page 271: A power function relates power to its determinants: (1) level of significance, (2) MDE size, (3) the unexplained variance of the outcome of interest, (4) allocation fractions, (5) and the sample size.

PRINCIPLE/ABBREV. ITEM	QUESTION	SCORE	DESCRIPTOR	SOURCE
Robustness of Methodology: Described sampling approach to collect qualitative data	[20] For the **qualitative** research methods used, is the sampling approach described in sufficient detail? (at a minimum, a rationale for the sample size and method of sample selection) and is it appropriate for the study objectives? **[Perf. Qualitative]**	yes/partial/no /NA	Researchers/evaluators should provide a description of the sampling frame and potential issues with it, if any. This should include an explanation of how the participants were selected, whether these participants were the most appropriate to provide access to the type of knowledge sought by the study, whether there was a point at which incoming data produced little or no new information (saturation) as well as any discussions around recruitment, such as why some people might have chosen not to take part in the study. "Partial" score should be given if only some of these elements were discussed. "NA" score should be given if this study did not use qualitative research methods.	CASP, Qualitative Checklist: Recommended considerations about "If the researcher has explained how the participants were selected"; "If they explained why the participants they selected were the most appropriate to provide access to the type of knowledge sought by the study"; If there are any discussions around recruitment (e.g. why some people chose not to take part)".
Robustness of Methodology: Conclusion	Robustness of methodology: Conclusion **[All evaluation types]**	adequate/not	**Adequate:** Overall, this evaluation demonstrates adherence to principles of appropriateness/rigor of chosen methodology **Not Adequate:** This evaluation contains major issues with the appropriateness of the chosen methodology, major deficiencies in the rigor with which it was applied or provides insufficient information for determining this	
Robustness of Methodology: Justification	Robustness of methodology: Notes/Justification **[All evaluation types]**		For instance: *"The study aims to identify and examine specific effects of receiving grants alone compared to receiving grants as well as training on student learning outcomes. The study clearly aims to establish a causal linkage between grants versus grants/training on student outcomes. The experimental design was, therefore, most appropriate to answer the research question. The study demonstrates rigorous application of the experimental technique within The Gambian setting. The authors clearly describe the interventions and adopt all the rigors of a well-applied randomization."*	
Validity: Addressed construct validity of the assessment tools	[21] Do indicators used in the evaluation capture the construct or phenomenon being investigated? **[IE, Perf. Quantitative]**	yes/partial/no	In order to assess the validity of the measurement, it is important to consider whether or not the chosen indicators adequately capture the concepts being measured or whether there are other dimensions central to the concepts that are being ignored, such as a labor market condition index that ignores underemployment. "Partial" scores could be given if some, but not all key indicators, adequately captured the concepts being measured.	BE2, page 24: In the case of measurement validity, it is important to repeatedly consider whether or not the indicator chosen fully captures the concept being measured. Are there other dimensions of the central concept that are being ignored?

PRINCIPLE/ABBREV. ITEM	QUESTION	SCORE	DESCRIPTOR	SOURCE
Validity: Addressed the external validity of findings from the sample to population	[22] Were the sampling conducted in such a way such that the results are generalizable to the population of beneficiaries reached through the activity? **[IE, Perf. Quantitative]**	yes/partial/no /NA	A number of characteristics of the survey design, such as timing of the assessment and absence of sampling weights, may affect the interpretation and/or calculation of population estimates. The evaluator/research may provide information about the timing of the assessment (e.g., pre-test and post-test being conducted at comparable time points in a cross-sectional design) or construction and use of sampling weights in the analysis (when different observations in a random selection process may have different probabilities of selection). "Partial" score could be given if the report mentions that the interpretation and/or calculation of some but not all population estimates took into account relevant survey design characteristics. "NA" score should be given in case this is a qualitative study.	StataCorp's Survey Data Reference Manual, page 3: In sample surveys, observations are selected through a random process, but different observations may have different probabilities of selection. Weights are equal to (or proportional to) the inverse of the probability of being sampled. Various postsampling adjustments to the weights are sometimes made, as well. A weight of wj for the jth observation means, roughly speaking, that the jth observation represents wj elements in the population from which the sample was drawn. Omitting weights from the analysis results in estimates that may be biased, sometimes seriously so.
Validity: Addressed ecological validity of findings	[23] Does the report allude to whether the study findings may have been biased by the activity of doing the study itself? **[All evaluation types]**	yes/no	Evaluators/researchers might discuss in the report whether findings could have been influenced by the process of research itself (ecological validity) or whether participants may have changed their behavior in response to their perception of the evaluators' objective (response bias), such as when the treatment group works harder than normal in response to being part of an evaluation (Hawthorne effects). Note that the tendency of participants to give an answer to a question that is in line with social norms even if this does not accurately reflect their experience (social desirability bias) is not relevant for this question. This might include discussions about whether the implementer may have brought in irreproducible energies that accountable for the success of a pilot but that might be absent in a scale-up.	BE[2], page 25: whether the findings could have been influenced by the process of research itself (ecological validity).

PRINCIPLE/ABBREV. ITEM	QUESTION	SCORE	DESCRIPTOR	SOURCE
Validity: Addressed the external validity of findings to other contexts	[24] Does the report address the external validity of findings? **[All evaluation types]**	yes/partial/no /NA	Findings are externally valid when they are valid in contexts other than those the evaluation was conducted in. Thus, researchers/evaluators may discuss the local conditions that would make it replicable in a different context. "Partial" score could be given if the external validity of some, but not all key findings, are discussed in the report. "NA" score should be given in case this evaluation did not intend to have data from a sample extrapolated to a population.	BE2, Checklist: To what extent is the study externally valid?
Validity: Data collection tools piloted with representatives of target populations	[25] Were all data collection tools piloted with representatives of target populations prior to beginning of the data collection? **[All evaluation types]**	yes/partial/no	Researchers/evaluators should describe if respondents used to pilot the data collection tools were similar to the target population of the full study. "Partial" score could be given if the report mentions that piloting was done but not with who.	EGRA Toolkit, 2nd edition, page 92: The students and schools selected for the pilot sample should be similar to the target population of the full study.
Validity: Confidence intervals reported around point estimates	[26] Are confidence intervals reported around point estimates? **[IE, Perf. Quantitative]**	yes/no/NA	USAID recommends that the margin of error be reported along with the findings from statistical samples. "NA" score should be given if the study does not use inferential statistical methods.	ADS 201sae: Has the margin of error been reported along with the data? (Only applicable to results obtained through statistical samples.)
Validity: Treatment effects presented in terms of effect sizes	[27] Is treatment effect presented in terms of effect size? **[IE, Perf. Quantitative]**	yes/no/NA	Researchers/evaluators often record the study findings in the units of the outcome variable. To improve the comparability of effect size estimates across outcome variables and across studies, effect sizes in terms of standard deviations should also be provided, taking into consideration the study design. "NA" should be given if the study did not conduct statistical hypothesis testing (as in the case of qualitative studies).	What Works Clearinghouse Procedures and Standards, page 22: For all studies, the WWC records the study findings in the units reported by the study authors. In addition, the WWC computes and records the effect size associated with study findings on relevant outcome measures.
Validity: Conclusion	Validity: Conclusion **[All evaluation types]**	adequate/not	**Adequate:** Overall, this evaluation demonstrates adherence to principles of validity. **Not Adequate:** This evaluation contains major deficiencies in establishing the measurement, internal, external or ecological validity or provides insufficient information for determining this.	

PRINCIPLE/ABBREV. ITEM	QUESTION	SCORE	DESCRIPTOR	SOURCE
Validity: Justification	Validity: Notes/Justification **[All evaluation types]**		For instance: *"The authors describe steps they took to address the validity of the study. For example, items included in the test had to relate directly to what grade 5 children would be expected to know at the start and end of the school year and statistical analyses were conducted to assess the internal consistency of questions in order to refine and adjust the assessment tools (measurement validity). In assessing learning progress of pupils in grade 5, the study included initial test scores into the estimation and controlled for background factors that may generate biases (internal validity). The study is based on longitudinal data collected from 5 provinces out of 58 in Vietnam, the generalizability of the findings is somewhat questionable (external validity), and there is no discussion of whether the findings could have been influenced by the process of research itself (ecological validity). While it could be improved, overall this study meets basic standards of scientific validity."*	
Reliability: Steps taken to ensure that data were reliably collected	[28] Does the report list steps taken to ensure that data were collected with a high degree of reliability? **[All evaluation types]**	yes/partial/no	USAID recommends that data collection methods be documented in writing to ensure that the same procedures are followed each time. The report may describe the use of data quality assurance checks such as accompaniments, back-checks and scrutiny, and these may have been conducted through spot-checking or for all questions in the data collection form. In case of paper-and-pencil data collection, double data entry report and/or double manual verification may also be mentioned in the report. Steps used in qualitative studies may include audio recording, videotaping and transcribing interviews. "Partial" score could be given if steps to ensure the reliability of some, but not all data collected, are described.	ADS 201sae: Are data collection and analysis methods documented in writing and being used to ensure the same procedures are followed each time?

PRINCIPLE/ABBREV. ITEM	QUESTION	SCORE	DESCRIPTOR	SOURCE
Reliability: Target and actual sample sizes reported and non-responses bias discussed	[29] Does the report adequately address missing data/non-response? **[All evaluation types]**	yes/partial/no	Researchers/evaluators should report the target number of respondents, the number of respondents reached, and the number of respondents who were included in the data analysis. This includes non-response in qualitative studies. For quantitative evaluations, the report may also mention using post-stratification to adjust weights for non-response. "Partial" score could be given if information about valid responses is provided to some, but not all data used in the findings.	What Works Clearinghouse Procedures and Standards, page D.4: study must report the number of students (teachers, schools, etc.) who were designated as treatment and comparison group samples and the proportion of the total sample (e.g., students, teachers, or schools in the treatment and comparison samples combined) with outcome data who were included in the impact analysis (i.e., response rates). Both overall attrition and attrition by treatment status must be reported.
Reliability: Conclusion	Reliability: Conclusion **[All evaluation types]**	adequate/not	**Adequate:** Overall, this evaluation demonstrates adherence to principles of reliability. **Not Adequate**: This evaluation contains major deficiencies in establishing the reliability of the measurement or provides insufficient information for determining this.	
Reliability: Justification	Reliability: Notes/Justification **[All evaluation types]**		For instance: *"This study used multiple researchers to undertake school observations and interviews; the researchers checked their own conclusions with each other and then cross-checked them against the wider analytical team to analyze between schools. The team ensured that different types of data were collected – observations, interviews and document analysis – to triangulate findings and take into account the variety of possible contexts. The authors also provide a good example of how to enhance the reliability of qualitative analysis: interviews were videotaped and transcribed."*	
Cogency: Answers to all study questions, including sub-questions, included	[30] Are all the study questions, including sub-questions, answered? **[All evaluation types]**	yes/no/NA	The purpose of an evaluation report is to provide the evaluators' findings and recommendations on each and every evaluation question. Accordingly, USAID expects that the answers to all evaluation questions, including any sub-questions, will be provided in the report. "NA" score could be given if no evaluation questions are provided in the report.	ADS 201mah: Address all evaluation questions in the Statement of Work (SOW) or document approval by USAID for not addressing an evaluation question.

PRINCIPLE/ABBREV. ITEM	QUESTION	SCORE	DESCRIPTOR	SOURCE
Cogency: Answers to all study questions included in the Executive Summary	[31] Does the Executive Summary include answers to all of the study questions? **[All evaluation types]**	yes/no	The executive summary must provide an accurate representation of the main elements of the evaluation report without adding any new material information or contradicting the evaluation report in any way. As such, it is recommended that all evaluation questions/issues, including any sub-questions/issues, will be provided in the Executive Summary.	ADS 201maa: The Executive Summary of an evaluation report should present a concise and accurate statement of the most critical elements of the report.
Cogency: Written in a language adequate to its stated audience	[32] Is the report accessible to the audiences for whom the report indicates it is written (e.g., minimizing technical jargon if intended to the general public)? **[All evaluation types]**	yes/no	Reports should be written in an accessible way to non-experts. Excessive use of research terminology is also undesirable; the report should favor terminology that its intended audience is expected to be familiar with.	USAID Evaluation Policy, page 10: USAID evaluations of all types will use sound social science methods and should include the following basic features: (...) Evaluation reports that are shared widely and in an accessible form with all partners and stakeholders, and with the general public.
Cogency: Connection between study questions, findings, conclusions and recommendations	[33] Are conclusions based on findings and are the findings related to the evaluation questions? **[All evaluation types]**	yes/partial/no	In order to strengthen the study's conclusion validity, USAID requires that evaluation findings be based on reliable quantitative and/or qualitative data, and that conclusions and recommendations should be based on these findings. USAID also encourages evaluators to present a clear progression from Study questions to Findings to Conclusions to Recommendations (if any) in their reports, such that none of a report's conclusions and recommendations appear to lack grounding. "Partial" score could be given if some supporting data is provided for some, but not all findings.	E3 Sectoral Synthesis Checklist, question 32: Can a reader can follow a transparent path from findings to conclusions to recommendations?
Cogency: Visuals are helpful for a non-technical audience to understand the findings	[34] Is the narrative in the report supported by charts, maps and infographics that help non-technical audiences easily understand the study findings? **[All evaluation types]**	yes/partial/no	Visuals must be used to facilitate understanding of the findings by general audiences. Visuals should be standalone, such that they are interpretable without the audience needing to read extra text. "Partial score" could be given if the report uses visuals to an insufficient extent.	EGRA Toolkit 2nd edition, page 120: Data visualization must be used to facilitate understanding of the findings by general audiences. Visualizations are "standalone," such that the visual is interpretable without the audience needing to read extra text.

PRINCIPLE/ABBREV. ITEM	QUESTION	SCORE	DESCRIPTOR	SOURCE
Cogency: Conclusion	Cogency: Conclusion **[All evaluation types]**	adequate/not	**Adequate:** Overall, this evaluation demonstrates adherence to principles of cogency. **Not Adequate:** This evaluation contains major deficiencies in demonstrating adherence to principles of cogency or provides insufficient information for determining this.	
Cogency: Justification	Cogency: Notes/Justification **[All evaluation types]**		For instance: *"The evaluation contains a clear, logical argumentative thread that runs through the entire report. This links the conceptual framework for the study to the data and analysis, and, in turn, to the conclusions. The conclusions are backed up by the evaluation findings."*	

References:

ADS 201 Additional Help: USAID Recommended Data Quality Assessment (DQA) Checklist. USAID, September 2016. https://www.usaid.gov/sites/default/files/documents/1870/201maa.pdf

Building Evidence in Education (BE_2) Steering Committee. *Assessing the Strength of Evidence in the Education Sector.* 2015. https://www.usaid.gov/sites/default/files/documents/1865/BE2_Guidance_Note_ASE.pdf . DFID.

General Accountability Office (GAO). *Foreign Assistance: Agencies Can Improve the Quality and Dissemination of Program Evaluations.* March 2017. http://www.gao.gov/assets/690/683157.pdf

Green, Andrew, and Sam Hargadine. *Sectoral Synthesis of FY2015 Evaluation Findings: Bureau for Economic Growth, Education, and Environment.* USAID, December 2016. http://pdf.usaid.gov/pdf_docs/PA00MP17.pdf

RTI International. *Early Grade Reading Assessment (EGRA) Toolkit: Second Edition*. USAID, March 2016. https://globalreadingnetwork.net/resources/early-grade-reading-assessment-egra-toolkit-second-edition

USAID Evaluation Report Requirements: A Mandatory Reference for ADS Chapter 201. USAID, September 7, 2016. https://www.usaid.gov/sites/default/files/documents/1870/201maa.pdf

USAID Evaluation Policy. USAID, October 2016. https://www.usaid.gov/sites/default/files/documents/1870/USAIDEvaluationPolicy.pdf

Qualitative Research Checklist. Critical Appraisal Skills Programme (CASP), March 13, 2017.

What Works Clearinghouse. *Procedures and Standards Handbook, Version 3.0*. 2014. https://ies.ed.gov/ncee/wwc/Docs/referenceresources/wwc_procedures_v3_0_standards_handbook.pdf