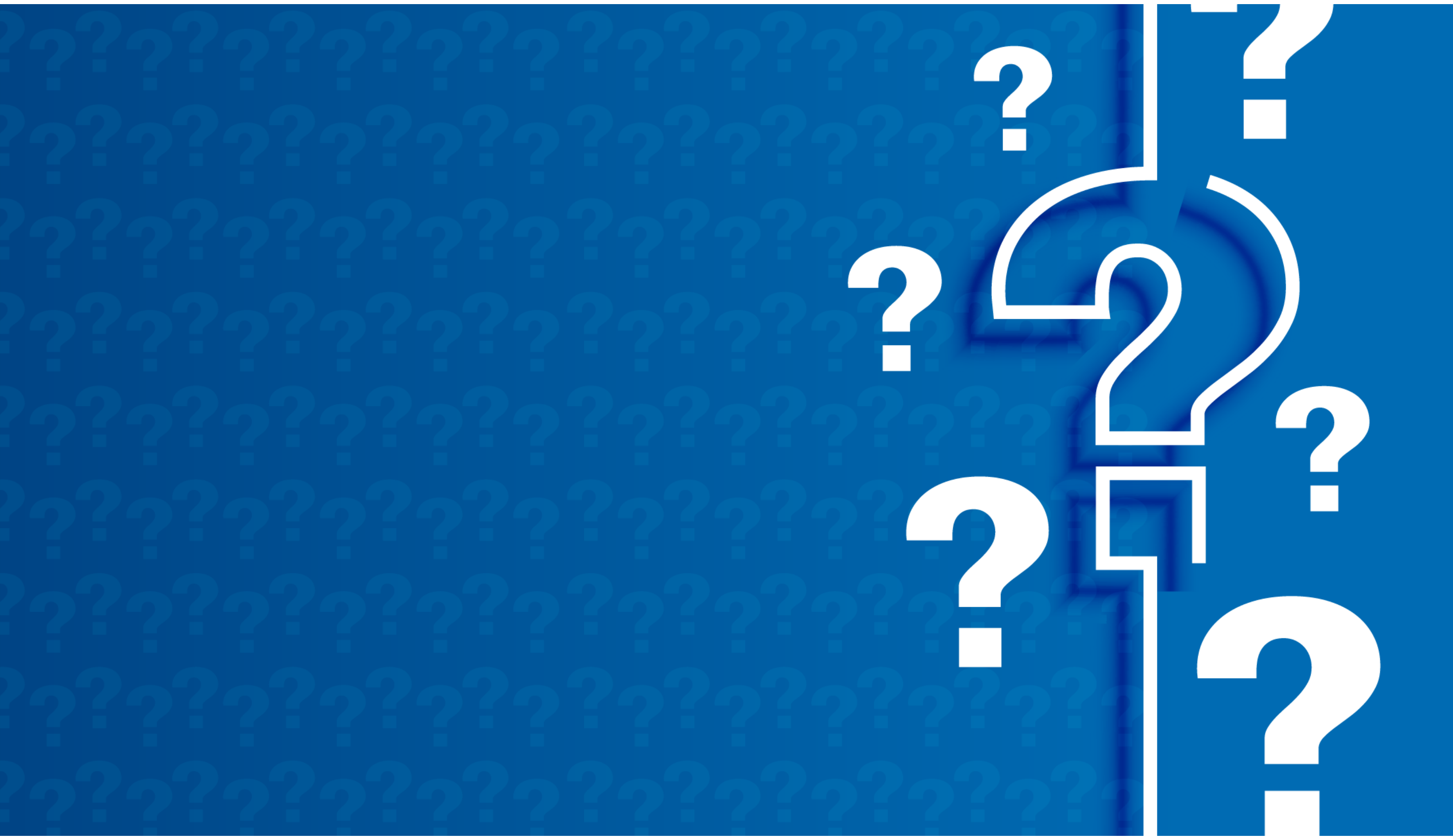


PHOTO: ADOBE STOCK

ADDRESSING LEARNING AND EVALUATION CHALLENGES—EVALUATION QUESTIONS

November 2024

Tasking CB072
Prepared under Contract No:
GS10F0218U/7200AA18M00017

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

This document was produced for review by the United States Agency for International Development, Democracy, Human Rights, and Governance Center under the Learning, Evaluation, and Research Activity II contract: GS10F0218U/7200AA18M00017. It was prepared by Andrew Green, Aleta Starosta, Ryan Hatano, and Daniel Sabet.

CONTENT

CONTENT iii

ACRONYMS iv

EXECUTIVE SUMMARY 1

BACKGROUND AND PURPOSE 2

OBJECTIVES AND RESEARCH QUESTIONS 3

APPROACH 3

REVIEW OF EXISTING GUIDANCE AND LITERATURE ON EQs 3

CREATION OF A PERFORMANCE EQ DATABASE 4

INTERVIEWS AND CONSULTATIONS 5

LIMITATIONS OF THE APPROACH 5

RQ1: CHALLENGES AND LIMITATIONS 6

CONCLUSION 1.1: THE MEDIAN NUMBER OF EQs EXCEEDS THE RECOMMENDED RANGE, BUT HIGHER QUESTION NUMBERS ARE NOT LINKED TO POORER-QUALITY EQs 6

CONCLUSION 1.2: OUTCOME QUESTIONS WERE MORE COMMON, BUT PROCESS QUESTIONS WERE OF HIGHER QUALITY 8

CONCLUSION 1.3: SCOPE, CLARITY, AND FEASIBILITY REMAIN PROBLEMATIC CHARACTERISTICS OF EQs 9

RQ2: ELEMENTS OF SUCCESSFUL QUESTIONS 11

ELEMENT 1: FEASIBILITY—CAN THE QUESTION BE ANSWERED WITH THE METHODS AVAILABLE? 11

ELEMENT 2: SCOPE—ARE PROJECT ACTIVITIES SPECIFIED IN REASONABLE QUANTITIES/TIMEFRAMES? 14

ELEMENT 3: CLARITY—IS THE QUESTION PRECISE? 15

PULLING IT ALL TOGETHER: ILLUSTRATIVE QUESTIONS BY COMMON EVALUATION TOPICS 18

RQ3: PROCESS IMPROVEMENTS FOR EQ DEVELOPMENT 20

RECOMMENDATIONS FOR AN IMPROVED EQ DEVELOPMENT PROCESS 21

APPENDIX 1: SUPPLEMENTAL GRAPHS 27

APPENDIX 2: LIST OF LER PEs 28

APPENDIX 3: INTERVIEWEES 32

ACRONYMS

AOR	Agreement officer's representative
ALEC	Addressing Learning and Evaluation Challenges
COR	Contracting officer's representative
DRG	Democracy, Human Rights, and Governance
EQ	Evaluation question
IE	Impact evaluation
KII	Key informant interview
LER	Learning, Evaluation, and Research
MEL	Monitoring, evaluation, and learning
PE	Performance evaluation
SOW	Scope of work
USAID	United States Agency for International Development

EXECUTIVE SUMMARY

Many evaluations, assessments, and studies conducted under previous Learning, Evaluation, and Research (LER) mechanisms in the Democracy, Human Rights, and Governance (DRG) Bureau, and under United States Agency for International Development (USAID) mechanisms more generally, involve a team of researchers conducting a large number of key informant interviews (KIIs) and focus group discussions at one point in time over a three- to four-week period of in-country work. As noted in the Addressing Learning and Evaluation Challenges (ALEC) Report[1], "basic" performance evaluations (PEs) regularly produce complaints, including about the quality of the evaluation questions (EQs) themselves.

The purpose of this research is to systematically explore the common challenges of developing quality DRG PE questions and to develop guidance for USAID staff to maximize the utility of PE questions. EQs could be thought of as scarce and non-renewable resources; there are only so many questions that could conceivably be answered, and PEs are typically one-off activities. The EQ development process involves a diverse set of stakeholders with differing incentives and perspectives in a process that quickly runs into the challenge of resolving multiple, interrelated tensions. What that process looks like and how it plays out will play an outsized role in the quality of EQs.

This findings report is guided by three research questions:

1. What are common challenges or limitations in USAID DRG PE questions and what are their primary causes?
2. What are example questions and key elements of example questions that a) can be answered with relative confidence by evaluation teams and b) meet the decision-making needs of USAID staff and implementing partners?
3. How can monitoring, evaluation, and learning (MEL) specialists help their colleagues translate their evaluative learning needs into questions that a) can be answered with relative confidence by evaluation teams and b) meet the decision-making needs of USAID staff and implementing partners?

The team considered the process for how USAID/Evidence and Learning team staff and other MEL staff can aid colleagues in setting parameters for the evaluation, brainstorming ideas, paring down questions to maximize both usefulness and answerability, and refining questions. The research team analyzed the EQ development process and quality of EQs through a review of practitioner guidance, existing PE questions, and interviews of USAID personnel and MEL platform experts.

For the first research question on evaluation question limitations, the team analyzed the common challenges in PE questions by creating a database of 548 EQs from 64 PEs conducted under LER contracts between 2014-2023, and then scoring the natures and characteristics of each EQ. The research team codes questions as either process vs outcome questions, descriptive or comparative questions, and scored the questions based on three characteristics identified in the practitioner literature: scope, clarity, and feasibility.. Scoring of each characteristic for every EQ was based on a three-point ordinal scale, and the team followed standard inter-reliability tests and processes to ensure consistent coding. Note that the database included no PEs from the Monitoring, Evaluation and Learning Services II indefinite delivery/indefinite quantity or mission-level MEL platforms, each of which may have somewhat different EQ development processes than that of the DRG LER mechanisms. At the same time as the coding and scoring process, the team interviewed nine evaluation commissioners and MEL platform leaders to collect information about the EQ formation process; note that this was an opportunistic sample with high rates of non-response.

[1] L. CAMACHO, K. MARPLE L. CAMACHO, K. MARPLE-CANTRELL, D. SABET. (2024). ADDRESSING LEARNING AND EVALUATION CHALLENGES.

Including sub-questions, the typical PE asked eight questions, above the recommended limit of five. The team found quality concerns across the three characteristics of scope, clarity, and feasibility, which points to a guidance gap addressing EQ design directly. Outcome questions were more common, but process questions were of higher quality. Of particular concern were comparative outcome EQs, which scored notably lower than all others. Scope, clarity, and feasibility remain problematic characteristics of EQs. While the median number of EQs exceeded the recommended range, higher numbers of questions were not linked to poorer-quality EQs. Based on the KIIs, the team posits that quality issues are likely generated by a process marked by poor cooperation and communication between stakeholders, as well as the strong influence of those without backgrounds in evaluation design and methodology.

For the second research question on example questions and elements, the team proposed guidance on the "Three Keys to Getting Evaluation Questions Right: Feasibility, Scope, and Clarity" for improving the quality of EQs based on the results of the first research question, including illustrative EQs by common evaluation topics. The KIIs provide key insights in the dynamic process of EQ development: evaluation purpose drives question development; USAID's consensus culture creates risks to EQ quality; and while approaches vary by mission, technical teams tend to be the most influential for EQ development. In response to these insights, the team provides guidance questions, process flowcharts, and call-out boxes to address common tensions to help stakeholders improve the EQ development process. In addition, the team developed a glossary of commonly-used terms such as sustainability and partnership.

For the third research question on developing question development guidance, the team made recommendations for improving the EQ development process. One recommendation is a five-step process rooted in best practices identified through KII analysis and existing USAID resources, which focuses on how to get input and buy-in **EARLY** from a variety of stakeholders in various interaction formats. Another recommendation is the use of the complementary Evaluation Question Development Workbook to help commissioners self-facilitate this process. EQs are a scarce and non-renewable resource, and quality can only come from active communication and deliberate cooperation.

BACKGROUND AND PURPOSE

Many evaluations, assessments, and studies conducted under previous Democracy, Human Rights, and Governance (DRG) Learning, Evaluation, and Research (LER) mechanisms, and under United States Agency for International Development (USAID) mechanisms more generally, involve a team of researchers conducting a large number of key informant interviews (KIIs) and focus group discussions at one point in time over a three- to four-week period of in-country work. For example, what might be called "basic" performance evaluations (PEs) are the most common form of external evaluation employed by USAID in the DRG sector. Such studies regularly produce complaints from commissioners, evaluators, and implementers related to the accuracy, reliability, and usability of findings. As discussed in the broader Addressing Learning and Evaluation Challenges (ALEC) Report[2], one common pain point in PEs is the quality of the evaluation questions (EQs) themselves. In fact, three common complaints about the quality and usefulness of performance evaluations can be directly linked to three key characteristics of an evaluation question used in this study: feasibility, scope, and clarity.

[2] IBID.

FEASIBILITY

If the evaluation questions (EQs) aren't feasible, commissioners can't be confident in the answers.

SCOPE

If the scope is too broad, evaluators struggle to tell commissioners information they don't already know.

CLARITY

If questions lack clarity, reports struggle to deliver what commissioners are expecting.

OBJECTIVES AND RESEARCH QUESTIONS

The purpose of this research is to systematically explore the common challenges of developing quality DRG PE questions and to develop guidance for USAID staff to maximize the utility of PE questions. This findings report is guided by the following research questions:

1. What are common challenges or limitations in USAID DRG PE questions and what are their primary causes?
2. What are example questions and key elements of example questions that a) can be answered with relative confidence by evaluation teams and b) meet the decision-making needs of USAID staff and implementing partners?
3. How can MEL specialists help their colleagues translate their evaluative learning needs into questions that a) can be answered with relative confidence by evaluation teams and b) meet the decision-making needs of USAID staff and implementing partners?

In answering these questions, the team considers the process for how MEL staff can aid colleagues in setting parameters for the evaluation, brainstorming ideas, paring down questions to maximize both usefulness and answerability, and refining questions. The team also considers when question formulation would benefit from the input of third-party evaluators or the use of multi-phased taskings and how this should be done.

APPROACH

The research team analyzed the EQ development process and quality of EQs through a review of practitioner guidance, a coding of existing PE questions, and interviews with USAID personnel and MEL platform experts.

REVIEW OF EXISTING GUIDANCE AND LITERATURE ON EQs

The research team reviewed literature developed by evaluation practitioners found in USAID's Learning Lab portal, the American Evaluation Association publications, and industry experts' online content to identify common elements or principles of quality EQs.

There was no consensus in the literature at a detailed level about what principles or characteristics make up a "good" performance EQ, but at a more abstract level, there was general acknowledgment that tension exists between questions that are answerable empirically versus those that are useful managerially. This discussion includes issues like the scope of the evaluation, common understanding of key terms or concepts, allotted time for data collection and analysis, and the alignment between questions and methods available per budgets and data sources. The research team decided to adopt guidance and supplemental material from USAID's Learning Lab Evaluation Toolkit, particularly the 2015 "Tips for Developing Good Evaluation Questions,"[3] to provide the most relevant and concise principles: 1) an actual question, not a statement, that both relates to USAID activity and is evaluative; 2) *limited* in terms of the number of questions and the extent of the activity to be evaluated; 3) *clear* in terms of the meaning of key terms like "effective" or "sustainable"; 4) *researchable* in terms of whether the scope of work (SOW) envisioned appropriate methodologies and resources as well as measurable standards or criteria; and 5) *useful* in terms of being tied to the evaluation's stated purpose and involving stakeholders.

The team adapted these five principles to identify a small number of key characteristics that could be reviewed and scored. Principles 1 and 2 were synthesized because only evaluative questions should be in a set of EQs; i.e., questions asking for recommendations or for causal analysis were excluded. The research team had no priors on the relationship between the number of EQs and EQ quality, leaving the characteristic of the **scope** of the question in terms of referring not just to the USAID activity, but to a specific and reasonable portion of that activity. Principles 3 and 4 were retained in whole as the **clarity** and **feasibility** characteristics. Principle 5 was included in whole for the coding pilot, during which the research team concluded that coding required review beyond the EQs that was far too labor-intensive.

CREATION OF A PERFORMANCE EQ DATABASE

The ET analyzed the common challenges and limitations of PE questions by reviewing 636 individual EQs from 64 PEs, representing the population of PEs conducted on DRG LER contracts between 2014–2023 (see Appendix 1). The team first coded the nature of the question:

- Is it asking a process question about the implementation of the activity or an outcome question about results or achievements?
- Is it descriptive (i.e., asking a "what is" question) or comparative (i.e., asking a specific target, result, state of being, or benchmark-based "what should be" question)?[4]

During this coding process, the team determined that 88 of the 636 (14 percent) EQs were either causal questions, requests for recommendations, or other non-questions, leaving 548 EQs for coding and analysis. Only 17 EQs (<3 percent) were causal questions, supporting the research team's prior hypothesis that the frequency of such questions had dropped in recent years, likely the result of guidance on impact evaluations (IEs). The team excluded these 88 items from scoring on question quality but included them in descriptive statistics about the database as a whole.

The research team then used guidance from the Learning Lab portal to code how "good" the question was by identifying three PE question characteristics:

- **Scope—***Are USAID project activities to be evaluated specified in reasonable quantities or timeframes?* EQs referencing the overall project as a whole are less useful for missions because limited evaluation resources means that teams focus efforts on activities that are small or of far less interest for future programming. Similarly, EQs incorporating timeframes that extend back before the project launch not only require expending resources to research non-project activities, but also face serious recall bias challenges.
- **Clarity—***Is the question a single clear, precise sentence free of jargon with key terms defined in the SOW?* Important concepts like sustainability or effectiveness need more specification in order

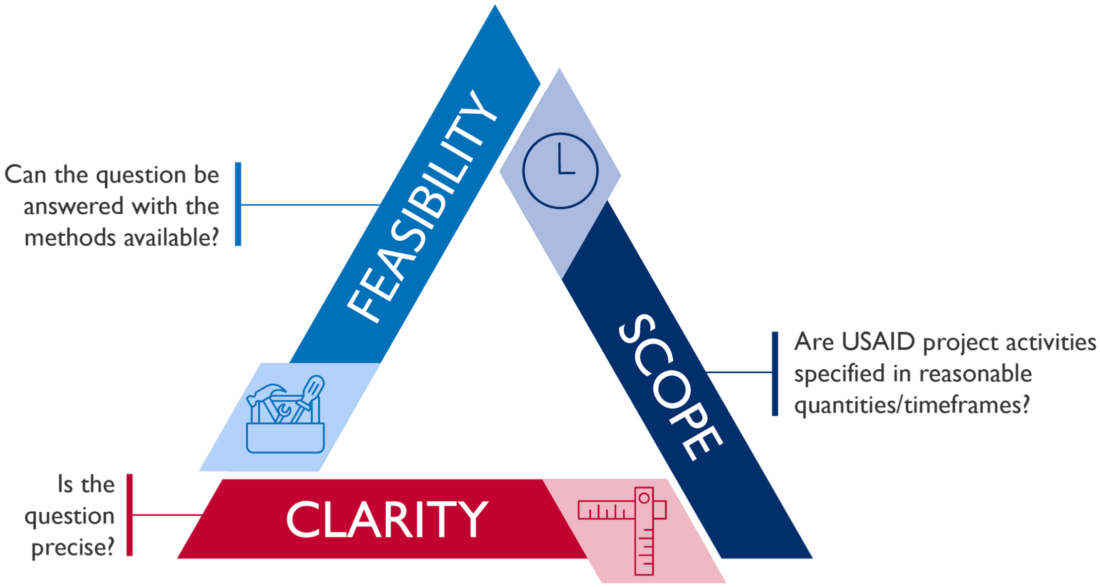
[3] Bureau for Policy , Planning, and Learning (2015) Tips for Developing Good Evaluation Questions. USAID.

[4] The literature (e.g., USAID's "Tips for Developing Good Evaluation Questions") typically uses "normative" where we use "comparative," as "normative" has a distinctly different use in political science and international relations.

for the evaluation team to collect appropriate data: sustainable in terms of financial resources or technical capacity? Effectiveness in terms of performance metrics or training outcomes?

- **Feasibility—***Are the evaluation's resources in terms of time, budget, available methodologies, and comparison standards or benchmarks appropriate for the EQ?* If the EQ requires a population survey to gauge citizen perceptions, the evaluation budget should include reasonable funds for that survey in addition to a timeframe that allows for good survey implementation practice. EQ language requesting comparisons must be sensitive to specifying the appropriate standards or benchmarks, as well as the cost and time of acquiring any externally produced data. Comparisons to other USAID or donor-funded projects, whether in the same country or elsewhere, can be highly problematic if those projects are not in fact obviously and readily comparable.

Figure I. Scope, Clarity, and Feasibility


Can the question be answered with the methods available?
FEASIBILITY
SCOPE
Are USAID project activities specified in reasonable quantities/timeframes?
Is the question precise?
CLARITY

The team used a three-value scoring system (i.e., fully present [2], partially present/absent [1], fully absent [0]) for each of the characteristics for the remaining 548 EQs. The research team added the ordinal scores for the three to create summary scores for each EQ (minimum total value 0, maximum total value 6) and then averaged the summary scores across the set of EQs for each PE. Note that the research team uses the summary scores of 0–6 when discussing EQs at the PE, database, or nature levels, but simplified scores of 0–2 when discussing EQs at the more detailed level of characteristics.

INTERVIEWS AND CONSULTATIONS

The team interviewed nine evaluation commissioners and MEL platform leaders to collect information about the EQ formation process. Participants included senior advisors and metrics specialists from USAID/DRG, representatives from USAID missions, senior MEL advisors from various organizations, and independent consultants with expertise in qualitative question design. The questions addressed issues in developing evaluation SOWs, improving the process of soliciting and refining EQs, and suggestions for enhancing existing guidance on writing EQs.

LIMITATIONS OF THE APPROACH

This research faced two limitations that could affect its findings and conclusions. The first is that the universe of cases was limited to those PEs that came through the LER mechanism, which often offered evaluation commissioners methodology feedback before SOW finalization. It is possible that PEs that came through evaluation-specific IDIQs, which tend to have limited or no feedback loops, or many mission-level MEL platforms, which have almost co-creation processes for evaluation SOWs, would generate different EQ scoring patterns. The second limitation is that the small number of interviews and high rates of non-response in a non-representative sample might have unintentionally excluded important perspectives.

RQ1: CHALLENGES AND LIMITATIONS

To answer the first research question on the common challenges or limitations in USAID DRG PE questions and their primary causes, the research team compiled and analyzed a dataset of EQs from LER PEs, augmented by qualitative interviews with evaluation practitioners.

Including sub-questions, the typical PE asked eight questions, above the recommended limit of five. The team found quality concerns across the three characteristics of scope, clarity, and feasibility, which points to a guidance gap addressing EQ design directly. Scope is particularly problematic, which has an indirect effect on feasibility. Outcome EQs, especially those incorporating comparison, have significantly lower quality than other EQs. Both quality issues most likely reflect the relatively stronger influence on EQ development by USAID professionals without evaluation design backgrounds.

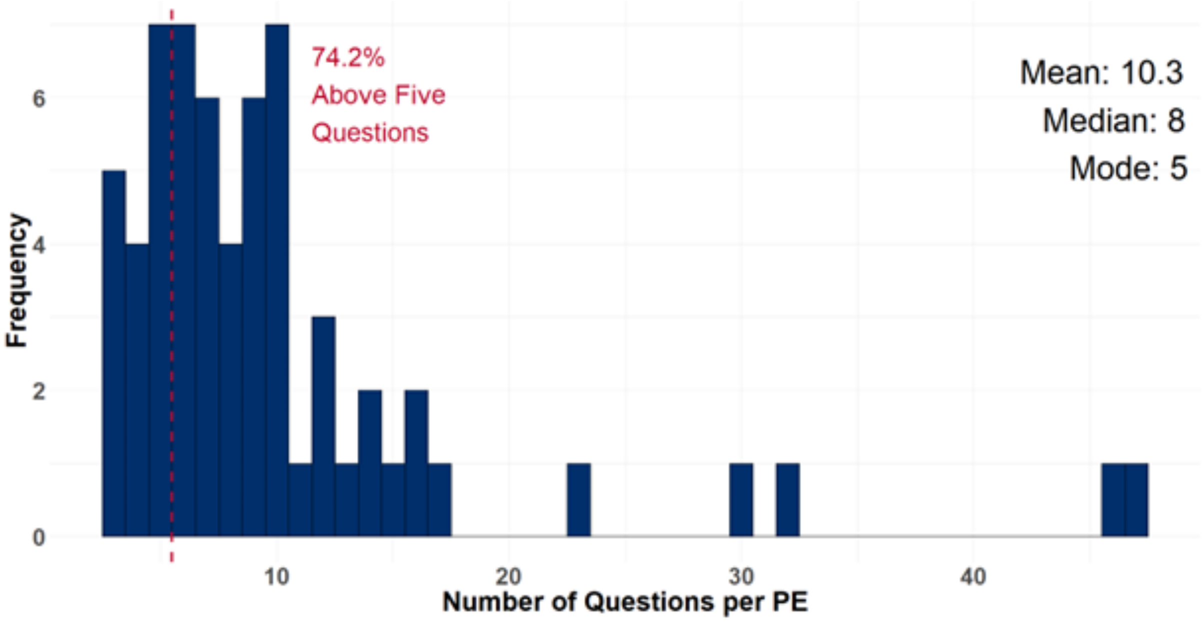
CONCLUSION 1.1: THE MEDIAN NUMBER OF EQs EXCEEDS THE RECOMMENDED RANGE, BUT HIGHER QUESTION NUMBERS ARE NOT LINKED TO POORER-QUALITY EQS

The "no more than five questions" guidance is known but acts more as a rule of thumb rather than a strict cap on the number of questions. The number of EQs does not appear to affect the quality of those EQs; rather, EQ quality is driven more by poor scope, clarity, and feasibility concerns.

The second principle in USAID's "Tips for Developing Good Evaluation Questions" resource specifically limits the number of EQs to no more than five. Based on qualitative interviews of experienced evaluation practitioners, it appears that the five-question limit is known and cited in evaluation development processes, even though the EQ development process discussed later in this report has dynamics that lead to more than five EQs in PE SOWs. Moreover, the interviewees were unanimous that issues of scope, clarity, and feasibility were still common.

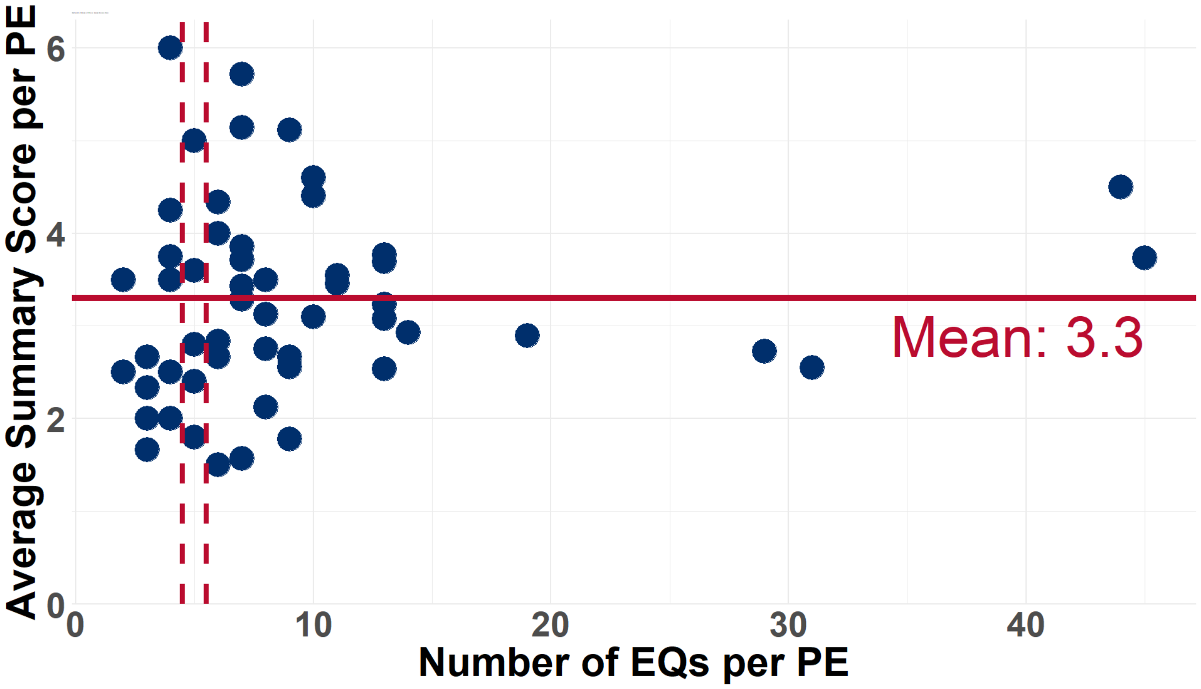
The research team found that the overwhelming majority of PEs had ten or fewer EQs when sub-questions were included. The overall distribution of the number of EQs in PEs was somewhat surprising, as the team found that the ***median*** number of EQs was eight, while the ***average*** of ten EQs was distorted by five evaluations with more than 20 EQs. The most frequent number of questions was five, six, and ten; 46 of the 64 PEs (74 percent) had more than five questions, but only 16 (25 percent) had more than ten EQs (see Figure 2).

Figure 2. Histogram of Number of EQs per PE



The research team found that there is likely no relationship between EQ quality and the number of EQs in a PE. The team found no pattern connecting the number of EQs in a PE and the quality of the set of EQs in a PE, which indicates that issues of EQ quality are not a function of the number of EQs. As Figure 2 shows, the average summary score by PE for the three characteristics of PEs with ten or fewer EQs varied from 1.5 to 6.0 for PEs with the most frequent number of EQs. The seeming randomness of EQ quality and number of EQs is likely due to the reality, shared by key informants, of different contexts and processes within each mission for the development of evaluation SOWs and EQs that lead to unclear evaluation utilization and overly broad scopes; a diverse body of stakeholders must grapple with tensions between answers and desire. Moreover, a "consensus culture" mentioned by key informants as a negative factor in the quality of evaluation design likely acts to prevent self-correction by actors with more knowledge of evaluation design.

Figure 3. Scatter Plot of the Number of EQs Per PE and the Quality of the Questions



CONCLUSION 1.2: OUTCOME QUESTIONS WERE MORE COMMON, BUT PROCESS QUESTIONS WERE OF HIGHER QUALITY

Outcome questions, particularly comparative outcome questions, present the biggest challenge to generating high-quality EQs, most likely as a result of the relatively stronger influence of technical teams on question design.

The research team found that outcome EQs dominated PEs, but they scored lower than process EQs on average, especially outcome-comparative EQs. As might be expected from a mix of midterm and final PEs, outcome questions dominated process questions (65 percent versus 35 percent). Descriptive questions and comparative questions occurred with almost equal frequency. As seen in Table 1 below and in Figures 6 and 7 in Appendix 1, the research team found that process EQs scored higher than outcome EQs, and descriptive EQs scored higher than comparative EQs; note the particularly low score of outcome-comparative EQs. Examples of outcome-comparative EQs from the dataset included EQs asking for a comparison to other programs, including some referencing programs by other donors, or to the perceptions of populations beyond those affected by the activity. Based on key informants' unanimity that the most influential actors in the development of EQs are technical teams (i.e., the mission personnel least likely to have any training in evaluation design), the team speculates that those actors are incorporating their wider technical knowledge into EQs that are beyond the scope or feasibility of an evaluation, reflecting the tensions between feasibility and what Missions want to know and between focus and breadth.

Table 1. Average EQ Scores by PE Nature on a Scale of 0 to 6

	PROCESS	OUTCOME	TOTAL
DESCRIPTIVE	**3.99** *101 (18%)*	**3.53** *179 (33%)*	**3.70** *280 (51%)*
COMPARATIVE	**3.23** *93 (17%)*	**2.75** *175 (32%)*	**2.92** *268 (49%)*
TOTAL	**3.62** *194 (35%)*	**3.15** *354 (65%)*	**3.31** *548 (100%)*

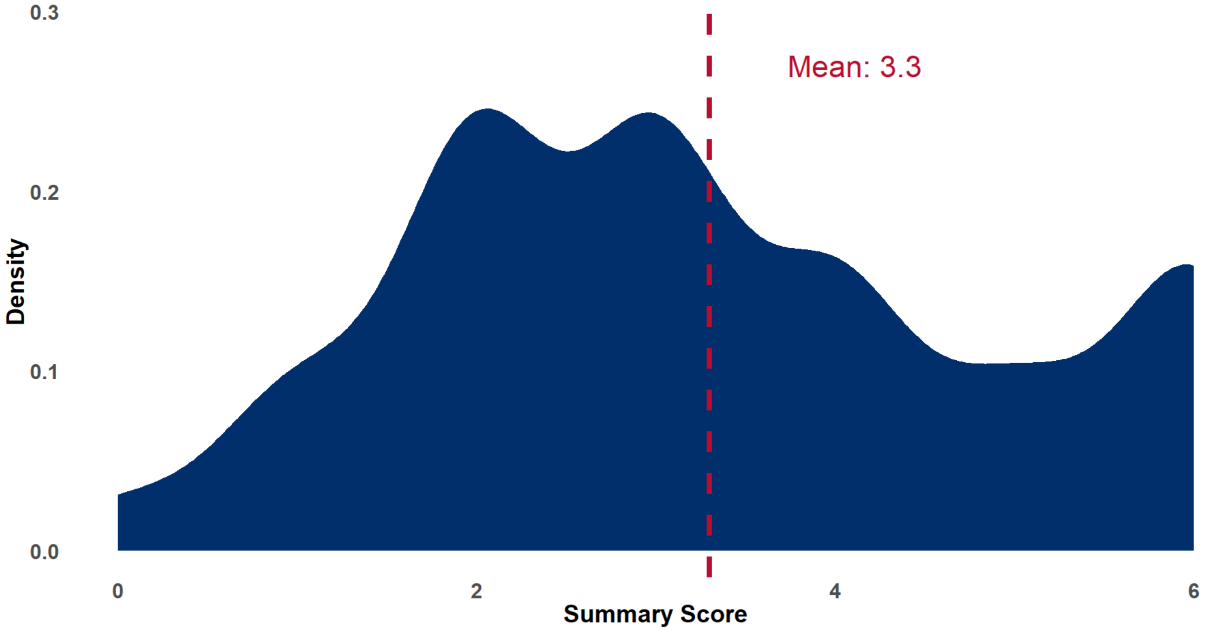
When reviewing only the highest-scoring EQs across the three characteristics, the research team found that outcome and particularly comparative EQs were significantly less represented than process or descriptive EQs. Further analysis of 143 EQs with summary scores of five or six (26 percent of the total EQs) found that each of the three characteristics earned the highest score in nearly equal numbers at the EQ level. The team found key differences in the disaggregation by process, outcome, descriptive, and comparative EQs, however. Of the 89 EQs (16 percent) with top scores for all three characteristics, there was a nearly equal split between process and outcome EQs (45 and 44 respectively), but a major gap between descriptive and comparative (68 and 21 respectively). Incorporating the 54 EQs that scored five overall (i.e., two out of three characteristics received top scores), revealed that the descriptive-comparative gap widened further to 103 versus 40 EQs. The Table 1 pattern is seen again when the research team considered these numbers in relation to the overall set: while approximately one-third of the process and descriptive EQs were high scorers, less than one-quarter of the outcome EQs and only 15 percent of the comparative EQs were. Outcome and comparative EQs are difficult to craft, and quality issues may be a manifestation of the influence of less knowledgeable actors in the EQ development process as they grappled with EQ design tensions.

CONCLUSION 1.3: SCOPE, CLARITY, AND FEASIBILITY REMAIN PROBLEMATIC CHARACTERISTICS OF EQS

Each of the three characteristics are problematic for EQ quality. Scope is particularly challenging, but clarity and feasibility are still not up to acceptable quality levels. This is most likely due to a misalignment of expectations between diverse stakeholders about what could be learned in a PE versus what is desired to be learned.

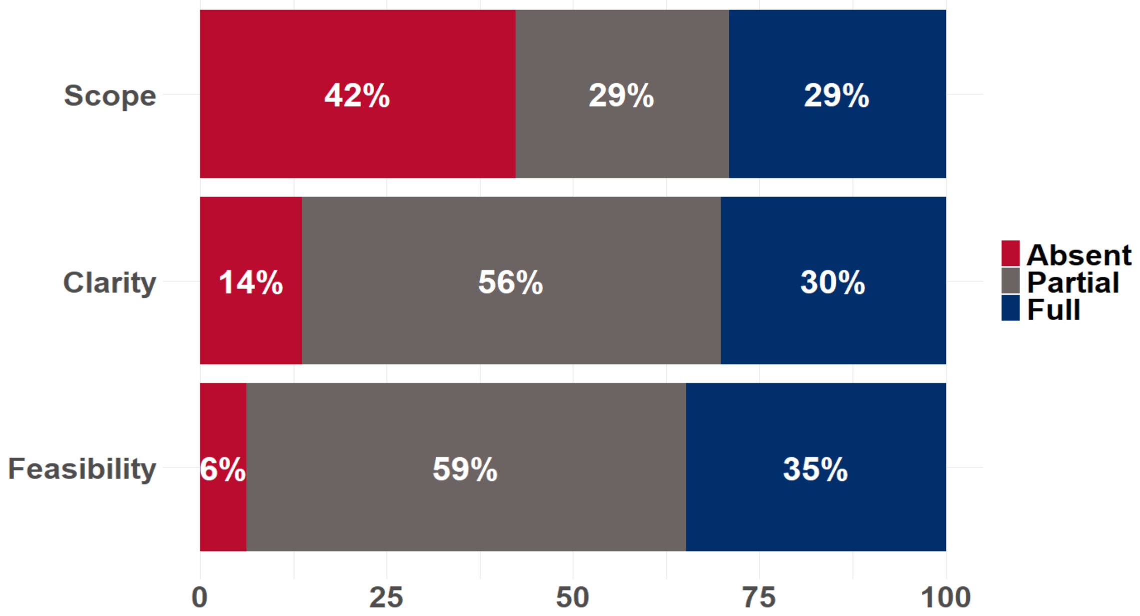
The research team found that the average PE summary score is equivalent to "partial" scores across the three characteristics. The overall average summary score for the set of 548 EQs was 1.1 on the 0 (fully absent) to 2 (fully present) scale (3.3 on the 0–6 scale), which roughly equates to "partial" scores on each of the three characteristics. Figure 4 shows further that a significant portion of PEs scored an average of two or three—many of the PEs averaged "partial" scores for the three characteristics. These data are corroborated by the qualitative interviews. Interviewees were unanimous that scope, clarity, and feasibility remained quality issues for EQs.

Figure 4. Density Plot of the Average EQ Summary Score



The research team found that only one-third of EQs met the full criteria for each of the three characteristics individually, and the plurality of scope scores was "fully absent." The team's analysis of characteristic-level scores produced multiple findings (see Figure 5 below). For each of the three characteristics individually, only approximately one-third of the EQs were of the highest quality; only 16 percent of EQs (89) had full scores for all three characteristics. Of the three, scope is clearly the most problematic, with a plurality of EQs scoring "fully absent." This overall picture comports with the results of the KIIs, in which participants unanimously pointed to the problem of different motivations, incentives, and knowledge bases across the typically diverse set of stakeholders. When the set of stakeholders expands beyond the COR and Program Office, evaluation utilization becomes unclear and viewpoints typically move up to the program or even multi-donor level, which, for example, leads to inappropriate scopes for evaluation.

Figure 5. Distribution of Average PE Scores by Scope, Clarity, and Feasibility



The research team found that clarity and feasibility were also problematic. The majority scores for clarity and feasibility were "partially absent," demonstrating that these are problematic characteristics for many EQs, too; note that the inclusion of the 17 causal questions removed from the set would not meaningfully change the score distribution for feasibility.

The specific problem for clarity is the well-known problem of unclear definitions of terms like "sustainability" or "effectiveness," a problem noted by all of the key informants. As seen in the dataset, the issues with feasibility were as much about a misalignment between the EQ and methods as about a poor understanding of the level of effort needed to collect required data. As speculated above, a lack of training in evaluation design could lead to the repetition of common question-wording errors like unclear definitions of terms or a misalignment between questions and methods.

RQ2: ELEMENTS OF SUCCESSFUL QUESTIONS

As previewed in RQ1, the key elements of a successful research question are scope, clarity, and feasibility. In this section, the research team proposes the following guidance for improving each element, based on findings from the research question coding, KIIs, and desk review.

ELEMENT 1: FEASIBILITY—CAN THE QUESTION BE ANSWERED WITH THE METHODS AVAILABLE?

There is a common tension between asking questions evaluators can feasibly answer with a PE methodology and asking questions that USAID needs answers to for upcoming programmatic decision-making. From one perspective, USAID should only ask performance EQs that can be answerable with the available methodologies, but from the other perspective, USAID staff should ask the questions for which they need answers. While both perspectives are valid, the more difficult the question is to answer with the available methods, the more likely that evaluation findings will have low levels of confidence and might lead to incorrect conclusions.

Traditional PEs are typically a combination of formative and summative evaluations. These evaluations are capable of answering questions about processes and outcomes, comparative questions, and questions about program assumptions and theories of change. They are not able to answer causal questions about program effects, which require an IE. If evaluation commissioners ask questions that cannot be feasibly addressed through a PE, the evaluation answers will have limited usefulness.

Question types that are feasible under most PE methodologies include:

- Questions about program **processes.**
- Questions about program **outcomes for program participants** (provided they can be measured with available methods and data - ask an evaluation specialist to confirm).
- Questions about program **assumptions and theories of change.**
- Questions that **compare program aspects.**
- Questions about **bottlenecks, challenges, and opportunities**.
- Questions about **lessons learned.**

The questions below are methodologically feasible but require data sources that go beyond the scope of most performance evaluations, which is addressed in the next section. Ask these questions with caution and ensure there are adequate resources to address them!

- Questions about outcomes for indirect participants or non-program participants.
- Questions comparing the evaluated program to other similar programs.[5]

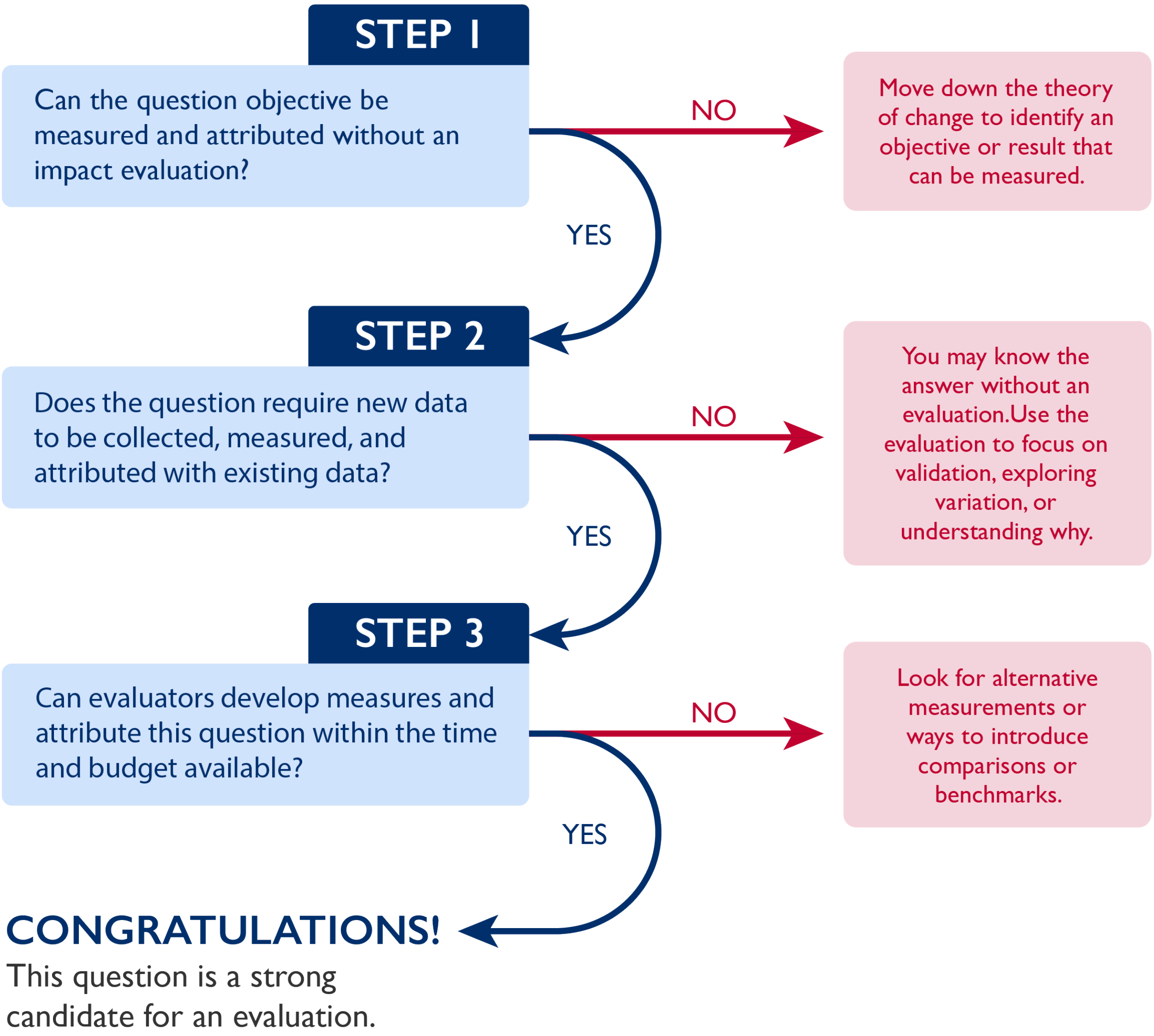
Ask these questions with caution and ensure there are adequate resources to address them!

Feasibility is particularly challenging when asking about effectiveness, results, or outcomes. In these cases, study commissioners should use the following flow chart to determine if a question is feasible.

[5] Comparisons must be made only after careful consideration of whether there actually is comparability. This is as true for site selection, as discussed in the ALEC Report, as it is for all small-n research. See Stanley Lieberson, Making It Count: The Improvement of Social Research and Theory (University of California Press, 1985); Charles Ragin and Howard Becker, What Is a Case: Exploring the Foundations of Social Inquiry (Cambridge University Press, 1992); and Gary King, Robert Keohane, and Sidney Verba, Designing Social Inquiry (Princeton University Press, 1994).

Figure 6. Feasibility Flowchart

To determine if a question is methodologically feasible:



STEP 1

Can the question objective be measured and attributed without an impact evaluation?

NO

Move down the theory of change to identify an objective or result that can be measured.

YES

STEP 2

Does the question require new data to be collected, measured, and attributed with existing data?

NO

You may know the answer without an evaluation.Use the evaluation to focus on validation, exploring variation, or understanding why.

YES

STEP 3

Can evaluators develop measures and attribute this question within the time and budget available?

NO

Look for alternative measurements or ways to introduce comparisons or benchmarks.

YES

CONGRATULATIONS!

This question is a strong candidate for an evaluation.

WHAT IF I WANT TO ASK A QUESTION THAT CAN'T BE ANSWERED?

With enough planning, money, and time, nearly all EQs can be feasible. **This is why evaluation planning needs to occur at the activity design phase.** If you have an impact question, plan for an IE at the beginning! If you want to know about changes in outcomes, then those outcomes should be measured through the activity MEL plan or a rigorous outcome PE. Without adequate planning, the best a typical qualitative PE methodology can do is provide is an impressionist answer that may or may not be valid. If a piece of information is critical to know, commissioners should consult with evaluation experts at the mission, bureau, or independent office to plan for an evaluation that can answer the question.

ELEMENT 2: SCOPE—ARE PROJECT ACTIVITIES SPECIFIED IN REASONABLE QUANTITIES/TIMEFRAMES?

USAID staff often want to understand the program as a whole or want the evaluation to speak to the many aspects of a program. This can lead to evaluation findings that are shallow and have limited confidence and limited usefulness. By focusing questions on **specific aspects** of a given program, theory of change, or implementation that are most important for informing decision-making, the evaluation team can look deeply into the most important issues and report their findings with confidence. To do this, commissioners should:

1. **Identify the decisions the evaluation will inform** and write questions aimed at providing this information. This goes beyond the broad purpose of the evaluation to specific design questions. For example:
   - Which program activities should be continued? Are there any gaps in current programming?
   - Is the program working with the correct stakeholders? What other partnerships should be explored?
   - Is the program reaching its target population? If not, how could targeting improve?
   - Is there uncertainty about any elements in the theory of change?
2. **Prioritize collecting information about the program components where there is most uncertainty**. For example, if MEL data provides clear information about the effectiveness of some activities, focus questions on other activities. If there is uncertainty about the theory of change, ask a question to validate the assumptions of a part of the theory of change.
3. **Questions that require in-depth comparisons to other donors, organizations, or programs can be valuable—but require more resources to answer.** If this information is critical for decision-making, ask the question, but ensure there are adequate resources to answer it by either reducing the number of other questions asked or adding resources to the evaluation.

4. **If asking a question about lessons learned, make sure it is tailored to specific activities or program components most relevant to future decision-making.** Avoid general questions about lessons learned, which frequently duplicate the findings and recommendations from other questions, and provide clarity on the types of lessons of interest. For example, ask about lessons learned in addressing a known and difficult programmatic challenge.
5. **Limiting each evaluation to 3–5 questions is a good rule of thumb—but the data sources and methods for answering each question matter more.** If multiple questions can be answered from the same data source (KIIs with the same stakeholders, a survey, MEL data, etc.), the ET can answer more questions. If a key question has a larger scope, such as comparisons to other programs or a question about the whole of the activity, even 1–2 questions may be all the ET can answer satisfactorily.
6. **If you must ask a question—or set of questions—with a large scope, recognize the trade-offs.** The evaluation team will have less confidence in the findings and the report will only go into limited depth. As such, it might not tell commissioners anything new. Increasing timelines and budgets can also make it possible to answer questions with a wider scope.

BUT WE WANT TO KNOW ABOUT EVERYTHING!

Every piece of an activity is important, and evaluation commissioners often claim they need to know about every part of a program. If the timeline and budget do not allow the ET to thoroughly investigate *all* components, prioritize based on the parts of the theory of change where there is the most uncertainty. For example, if a program included a training, did the training produce the intended behavior? If a particular service was offered, did the target population take it up? Going back to the intended purpose and use for the evaluation can help you prioritize which components of the program to focus the evaluation on.

ELEMENT 3: CLARITY—IS THE QUESTION PRECISE?

The third element of a successful question is **clarity**. If evaluation questions and terms used are not clear, evaluation teams might not provide the answers commissioners expect. To ensure questions are clear and precise, follow these guidelines.

1. **Express the question itself as a single clear, precise sentence free of jargon.** Provide additional context alongside the question, rather than as part of the question. Include additional context to guide the ET, including identifying lines of inquiry and potential hypotheses, providing examples, or giving additional information. This can be done in a paragraph placed under each EQ or included elsewhere in the evaluation SOW.

   For example, instead of this question:

   What have been the effects of GDP interventions (i.e., grants, trainings, research, communications labs, and economic empowerment work) on beneficiaries and their communities? In particular: What have been the most significant changes for groups within the LGBTQIA+ community (from the perspective of project beneficiaries and implementing/resource partners) as a result of the GDP?

Ask this:

What have been the effects of GDP interventions on beneficiaries and their communities, particularly the LGBTQIA+ community?

Interventions of interest include grants, trainings, research, communications labs, and economic empowerment work.

2. **Define all terms in the question.** What do you mean by resilient? Sustainable? Effective? Include clear definitions of any key terms used in the question. Use the glossary in Figure X below as a starting point.
3. **Use compound questions and subquestions sparingly and thoughtfully.** Evaluation commissioners will sometimes join two separate questions as one. This can make sense if the two components are asking for more detail about the same core question (such as "to what extent" or "why or why not") but in most cases, the two questions are better addressed separately. For example, the question below asks about participant perceptions, with a follow-on question about how the process could be improved.

Did those groups feel served by this process and, if not, what measures could be taken to improve upon the representation of marginalized groups, including women, in the peace process?

For clarity, this question should be simplified into three questions, but can be presented as single area of inquiry.

1. *Did marginalized groups, women in particular, feel served by the peace process? Why or why not?*
    a. *To what extent did the theory of change, program structure, and intervention design enhance or dampen the representation of marginalized groups in the peace process?*

The inclusion of "Why or why not?" strengthens the first part of the question, changing it from a yes or no answer to an open-ended question. This is an appropriate use of a compound question. The second part of the question should be pulled out as its own unique question, but it is an appropriate use of a subquestion. Because the answer to the second part of the question relies on the same data sources to answer, and the findings fall logically under the same EQ heading, it qualifies as a subquestion, rather than a separate question.

In contrast, the questions below (presented as a single question in the evaluation SOW), asks five questions on four distinct topics.

To what extent have the deliverables and outputs set forth in the task order and work plans been met? What factors have affected the project's success, including but not limited to relationships among key stakeholders? What lessons regarding the integration of key stakeholder interests can inform future USAID programming? What specific project achievements are or are not sustainable, and what is required to ensure sustainability?

A more clear way to present these questions is to separate out each question, narrow the scope of the second component to focus exclusively on key stakeholders, and make the third component on sustainability a more concise and informative question by comparing program elements to one another and focusing on the mechanism behind the sustainability of some achievements over others.

1. *To what extent have the deliverables and outputs set forth in the task order and work plans been met?*
2. *How and to what extent have relationships among key stakeholders affected the project's success?*
    a. *What lessons regarding the integration of key stakeholder interests can inform future USAID programming?*
3. *Which project achievements are most likely to be sustainable and why?*

Figure 7. Glossary of Evaluation Terms

Effectiveness: Successful in producing a specific desired result. Results may include program intermediate results or outcomes, but the former is more feasible to determine. Consider using defined categories (very effective, effective, less effective, not effective) to help make the question more clear.

Sustainability: The ability to maintain a specific outcome beyond a defined period of time, such as the program period of performance. Specify the outcome of interest (partnerships, financing, systems or processes, knowledge management, etc), as well as the time period.

Inclusivity: Ensures all people can participate in and benefit from USAID's development efforts. Specify any groups of particular relevance of interest, potentially including women, youth, LGBTIA+, and religious or ethnic minorities. For example, specify "impacts on women" instead of "impacts on disadvantaged groups."

Capacity: The ability of people, organizations, or networks to take action to solve local development challenges, learn and adapt, and innovate. Specify the type of capacity of interest—this may be related to knowledge, skills, motivations, and relationships.

Partnerships and stakeholders: Stakeholders are individuals, groups, or organizations that can positively or negatively impact the program outcomes. Be sure to specify the type of stakeholders or partnerships that are most relevant.

Resilience: The ability of people, households, communities, countries, and systems to mitigate, adapt to, and recover from shocks and stresses in a manner that reduces chronic vulnerability and facilitates inclusive growth. Be sure to specify who is expected to experience changes in resilience and what these changes are likely to be.

Adaptive management: An approach to implementing the program cycle that seeks to better achieve desired results and impacts through the systematic, iterative, and planned use of emergent knowledge and learning throughout the implementation of strategies, programs, and projects. Be sure to specify both the adaptive management strategies and the results and impacts of interest.

PULLING IT ALL TOGETHER: ILLUSTRATIVE QUESTIONS BY COMMON EVALUATION TOPICS

TOPIC	ILLUSTRATIVE QUESTION
Outcomes, effectiveness, and program objectives	Were the specific approaches developed under Objective 1 very effective, effective, less effective, or not effective at producing engaging media content? *Focusing scope through "specific approaches developed under Objective 1," while clarifying measurement terms with four rating categories and defining "effective" as "producing engaging media content." The term "engaging" should be defined elsewhere.* What have been the bottlenecks and opportunities in the activity approach to building legitimate, trustworthy, and responsive relationships between police and communities? *Focusing scope through "approach to building […] relationships between police and communities," and providing clarity elsewhere with definitions of "legitimacy, trustworthy, and responsive."* Which collaborative activities were valued most by journalists participating in the activity? Which of the three program approaches appear to be the most promising at enabling participating children to live in family care? Which program approaches under Objective 3 may have contributed the most toward increased service delivery? To what extent, if any, has the activity contributed to increasing the capacity of local civil society organizations to advocate for national policy changes? To what extent has the activity contributed to the institutional capacities of key justice sector institutions to address judicial corruption?
Sustainability	How likely is it that the activity's local government training and support program is able to be sustainably implemented by the national government one to three years after the end of the program? *Focusing scope through "local government training and support program," defining sustainability in terms of the national government implementing the program "one to three years after the end of the program." Note that for reasons of feasibility, this question is not asking for comparison to another donor's program.* What steps is the activity currently taking to build sustainability into its intervention and to what extent do these steps address existing constraints to sustainability?

TOPIC	ILLUSTRATIVE QUESTION
Theory of change and assumptions	To what extent did the mismatch between elements of the theory of change and the distribution of project human/financial resources affect implementation? *Focusing scope of factors affecting implementation to "the distribution of project human/financial resources," and enhancing feasibility by pointing to existing project data instead of e.g. a population survey.* Did the program have any unintended negative consquences on women's political participation? In what ways did changes in the context of political party competition and independent media capacity assumed by the theory of change generate obstacles or opportunities for the main activities under evaluation?
Inclusivity/gender and targeting	To what extent were women and youth included in community outreach about participatory budgeting? *Focusing scope on "women and youth" and "community outreach about participatory budgeting," and enhancing feasibility by pointing to information available from project MEL data and interviews of project staff.* To what extent and how did USAID activities foster the participation of target groups in civic education opportunities? To what extent did the project's prioritization of marginalized groups align with the expectations of key stakeholders?

TOPIC	ILLUSTRATIVE QUESTION
Implementation and adaptation	**To what extent and h**ow did the Mission use research and analysis, including previous evaluation findings, to make timely and effective programmatic changes to achieve its goals? To what extent is MEL data providing actionable information that is used to inform program adaptations? *Focusing scope on "program adaptations" and "MEL data," clarifying "actionable" elsewhere, and enhancing feasibility by pointing to information available from project MEL data and interviews of project staff.* What lessons should USAID and its partners draw from the first half of the activity implementation to inform adaptive management? To what extent have scale-up efforts incorporated lessons learned from the initial pilot? To what extent have lessons learned been identified, shared and incorporated across sites and sub-awardees?
Partnerships and stakeholder engagement	To what extent, if any, has the activity's existing partnership structure with the Ministry of Community Development and the Ministry of Finance contributed to the success or failure of the program to enable the sub-national governance system to expand revenue? To what extent is the activity engaging with the most relevant institutional stakeholders, including teachers, school boards, and government officials, at the provincial and national levels? *Focusing scope on "Objective 2 activities" and "most relevant institutional stakeholders [...] at the provincial and national levels, clarifying definitions of "engaging" and "most relevant" elsewhere, and enhancing feasibility by pointing to population set of "teachers, school boards, and government officials."*

RQ3: PROCESS IMPROVEMENTS FOR EQ DEVELOPMENT

The ET's KIIs provided insights into the dynamic process of EQ development. The interviewees were unanimous on three general process points:

1. **The evaluation purpose drives question development (in theory)**

   Across all KIIs, respondents stressed the importance of grounding the EQs in the purpose of the evaluation. "The key is addressing utilization from the very beginning," stated one evaluator, and Mission and USAID/DRG staff shared similar sentiments. However, putting this into practice proves challenging, and Missions shared stories where technical teams struggled to articulate the purpose of the evaluation or get adequate buy-in from key stakeholders, such as the office director, about the purpose of the evaluation. It is critical that evaluation commissioners take the time to clearly articulate and agree on what the purpose and utilization goals of the evaluation findings are before starting to develop EQs. It is helpful to not just identify a broad purpose (e.g., inform a follow-on activity) but to identify specific decisions that need to be made for that purpose (e.g., changes to program components, theory of change, geographical targeting, target population, program scope, partnership strategy).

2. **USAID's consensus culture creates risks to question scope and feasibility**

   The ethos of soliciting buy-in or concurrence from a wide set of evaluation stakeholders results in increased question counts, undefined terms, and misalignment between questions and methods. Any one stakeholder generally will not share the same motivations, incentives, or knowledge base with all other stakeholders. For instance, to paraphrase one example from a KII, the front office may view the evaluation as a way to solicit Congress or the agency for additional DRG funding, the COR or AOR of the program may view the evaluation as a check-the-box exercise, and the office director may view it primarily as a way to get insights into the design of the next program. Creating clear, feasible questions with a realistic scope that serve all three utilization goals is unlikely. As noted by one KII respondent, "Each mission is its own evaluation context and set of stakeholders," and it can be a struggle to limit question scope and feasibility when these stakeholders have different ideas about the purpose of the evaluation, as well as different levels of methodological expertise. It is therefore important to address this tension early in the question development process and prioritize among differing goals.

3. **Every Mission context is different, but technical CORs/AORs tend to be the most influential actors in the question development process**

   No one mandated procedure or guidance document for developing EQs would ever work for every mission. Every mission has a different program office capacity, different relations between the program office and technical teams, different levels of engagement by the technical office or mission leadership, and different overall sets of internal stakeholders. However, it was clear from the unanimity of key informant opinion that the technical team, particularly the COR/AOR of the activity to be evaluated, is a key target for guidance on developing strong EQs.

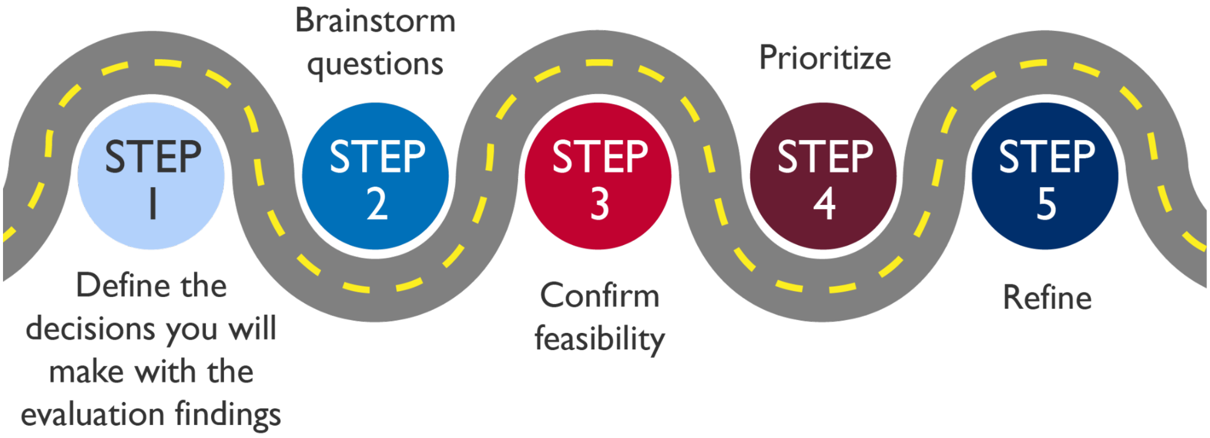
   According to USAID KIIs, "Technical teams need to be involved because they understand the activity," and "It is critical to have good rapport with the technical teams." Technical teams are also the most direct user of the evaluation findings. Technical teams bring the most detailed expertise about the project, but they may lack the evaluation expertise to understand the feasibility of certain types of questions, particularly causal questions. According to interviews with Mission staff, non-evaluation specialists generally play the strongest role in shaping the number and content of EQs. This is less true for large missions with commensurately larger program offices that could attract evaluation specialists with technical knowledge.

RECOMMENDATIONS FOR AN IMPROVED EQ DEVELOPMENT PROCESS

To improve the EQ development process, the research team recommends the following five-step process, rooted in best practices discovered during KIIs and documented in existing resources, such as USAID's Learning Lab's Evaluation Toolkit. This requires getting input and buy-in from a variety of stakeholders, including the activity COR/AOR, office leadership, and sometimes actors, including the implementing partner, host country government, or other donors. This process could be conducted in many different formats and incorporated into already existing processes for developing EQs. Discussions could be held in a single facilitated workshop, a series of more informal meetings, or over e-mail or through a shared document—the format is less important than ensuring this conversation happens early in the process.

The research team has developed an Evaluation Question Development Workbook to help commissioners self-facilitate this process. The process is outlined in Figure 8.

Figure 8. Performance Evaluation Question Development Process



STEP I: DEFINE THE DECISIONS TO BE MADE WITH THE EVALUATION FINDINGS

All successful evaluations start with a clear purpose. Commissioners invest time and money in conducting the evaluation, and taking time in the beginning to clearly define what decisions they plan to make based on the evaluation findings will help ensure the EQs produce information they can use.

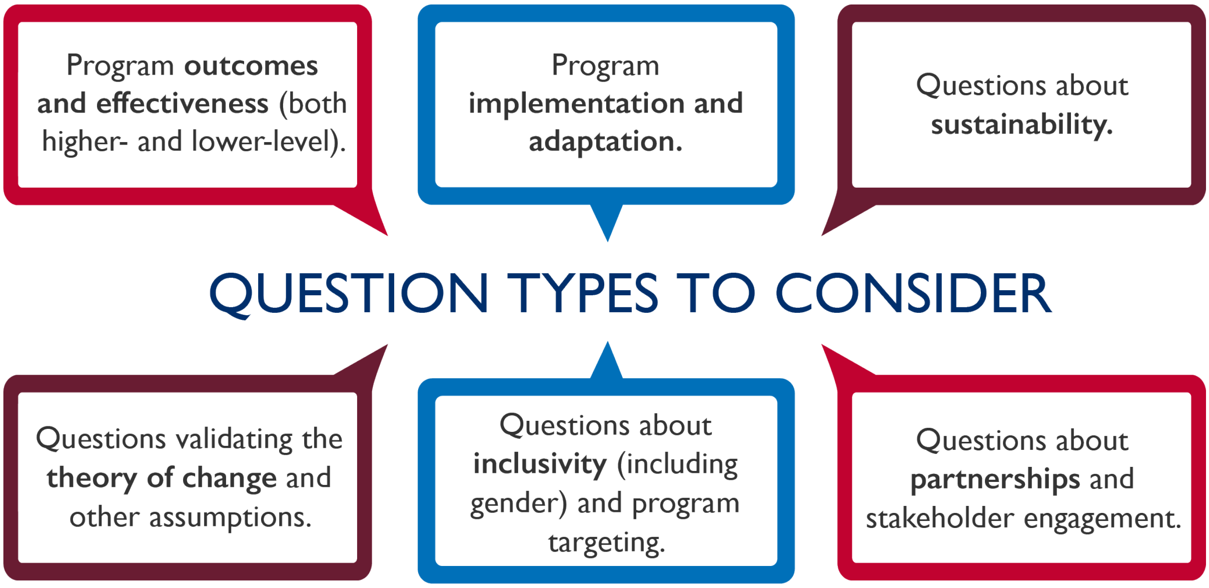
Some questions to consider include:

- What decisions do we need to make for a follow-on program? Examples: program components, theory of change, targeting, program scope, other U.S. Government funding paths/partners, or even division of labor amongst donors.
- What adaptations might be needed for the implementation? Examples: To increase inclusivity, increase sustainability, scale the program, shift intervention implementation from project staff to grantees.

STEP 2: BRAINSTORM QUESTIONS

When the purpose and the decisions the evaluation will inform are clear, conduct a brainstorming session to gather a comprehensive list of questions from all necessary stakeholders, ensuring that everyone's input is considered. This stage focuses on generating a wide array of questions without worrying about exact wording or prioritization.

Below are some common question types you may want to consider based on the decisions you plan to make:



QUESTION TYPES TO CONSIDER

Program **outcomes and effectiveness** (both higher- and lower-level).

Program **implementation and adaptation.**

Questions about **sustainability.**

Questions validating the **theory of change** and other assumptions.

Questions about **inclusivity** (including gender) and program targeting.

Questions about **partnerships** and stakeholder engagement.

STEP 3: CONFIRM FEASIBILITY

For each question, use the following chart to help determine if the question is methodologically feasible. This is the time to bring in specialized evaluation knowledge, either through a Mission MEL expert, a MEL platform, USAID/Washington, or a contractor. Use the flowchart shown in Figure 6 for additional guidance on questions related to effectiveness, results, and outcomes.

STEP 4: PRIORITIZE QUESTIONS

Once the universe of questions has been narrowed to those that are methodologically feasible, commissioners will need to prioritize questions until they reach a single question set that can be answered with the time and money available. One tool to help prioritize questions is a variation of the popular management prioritization framework, RICE (Reach, Impact, Confidence, and Effort).

REACH

What will this question help us learn?

IMPACT

How much will this information impact our decision-making?

CONFIDENCE

How confident are we in the quality of the findings?

EFFORT

What types of data are necessary to answer this question, and does the evaluation have the time and money to collect it?

In most cases, this will be primarily a qualitative exercise to think about the tradeoffs of various questions. However, commissions could also provide a quantitative score to each category by creating a three-point scoring system and multiplying reach by impact and confidence, then dividing by effort. When prioritizing, be sure to consider how the questions work together as a whole and be mindful of the resources available to answer them. A question with a large scope—such as a comparison to other programs or a question about the effectiveness of all the components of an activity—may be worth asking if it is the information that will most inform the decisions commissioners need to make, but that may be the only question the evaluation can answer. Similarly, if a question can only be answered by the methods available with a limited amount of confidence, it may not be the best use of resources to ask it. Commissioners may wish to get additional inputs from mission, bureau, or independent office MEL specialists, learning experts in Washington, or the ET conducting the evaluation at this stage. They will be well positioned to provide guidance on confidence and effort.

STEP 5: REFINE QUESTION SCOPE AND CLARITY

Once you have narrowed your list to your highest-priority questions (no more than five is a good rule of thumb) use the checklist to make further refinements to the scope and clarity of the question. Ensure all project activities are specified in reasonable qualities and timeframes, use a paragraph to provide context, define terms, share additional information or nuance, and identify lines of inquiry. You may also wish to consult with MEL specialists at the Mission or MEL platform, learning experts in Washington, or the ET conducting the evaluation to help refine the questions. Once you have refined your questions, circulate them for any necessary approvals. If your evaluation partner has not yet seen the questions, it can be helpful to note whether your team is open to suggestions from the ET to further refine your questions during the work plan stage, or if the question set is unchangeable once approved.

APPENDIX I: CODING METHODOLOGY

The team analyzed the common challenges in PE questions by creating a database of 548 EQs from 64 PEs and then coding a set of variables for each EQ. Standard processes of inter-rater reliability analysis, item-level review and discussion, consensus, and guidance review were followed.

Data Collection

The 64 PEs came from the set conducted on DRG LER contracts between 2014–2023, and sentences were scraped from the Evaluation Questions section of PE reports. After cleaning the initial dataset to include only sentences ending in a question mark, the resulting dataset contained 636 items; note that this dataset including questions and sub-questions. The initial coding process identified a further 88 items that were either causal questions, requests for recommendations, or some other non-question, leaving 548 EQs for coding and analysis. The items were left in the dataset for descriptive statistics purposes, but excluded from the coding process.

Coding EQs

The basic coding process after removing the 88 items was to code each question for variables of nature and characteristic. Coding was piloted on 36 EQs, inter-rater differences were analyzed statistically, and then reconciliation of each difference was made through an internal workshop. Inter-rater reliability analysis was later conducted on the complete dataset, and differences reconciled in an internal workshop.

The natures coding captured process vs outcome and descriptive vs comparative through binary values of present [1] or absent [0].[6]

- Process vs outcome – Is it asking a *process* question about the implementation of the activity or an *outcome* question about results or achievements?)
- Descriptive vs comparative – Is it *descriptive* (i.e., asking a "what is" question) or *comparative* (i.e., asking a specific target, result, state of being, or benchmark-based "what could or should have been" question)?)

The characteristics of scope, clarity, and feasibility were identified through practitioner literature review.[7] Scoring of each characteristic for every EQ was based on a three-point ordinal scale of fully present [2], partially present/absent [1], fully absent [0].[8]

- Scope—Are USAID project activities to be evaluated specified in reasonable quantities or timeframes?
- Clarity—Is the question a single clear, precise sentence free of jargon with key terms defined in the SOW?
- Feasibility—Are the evaluation's resources in terms of time, budget, available methodologies, and comparison standards or benchmarks appropriate for the EQ?

[6] The team attempted to use ChatGPT to code for each of these natures, but the error rate was quite high.

[7] The team also attempted to capture a "Relevance" characteristic, i.e., how relevant the EQ was to the stated evaluation purpose, but the high level of effort to capture that information and then code led to its exclusion from this exercise.

[8] The team first applied binary coding to the characteristics, but decided at the end of piloting to adopt a three-value ordinal scale, which resulted in repeating the coding and piloting analysis.

The team followed standard inter-reliability tests and processes to ensure consistent coding in piloting and overall phases.

The team piloted coding by three raters for natures and characteristics on a set of 36 EQs from three PEs. Inter-rater reliability analysis and EQ-level review showed minor differences in coding, which were resolved through an internal workshop and adjustments in coding guidance.

The team of three raters coded the entire dataset and again conducted a process of inter-rater reliability analysis and EQ-level review. Analysis showed systematic differences for each of two raters for different characteristics, but the differences were largely limited to a single ordinal value. An internal workshop to review differences at the EQ-level produced consensus on coding.

APPENDIX 2: SUPPLEMENTAL GRAPHS

Figure 7. Density Plot Comparing Process and Outcome Questions

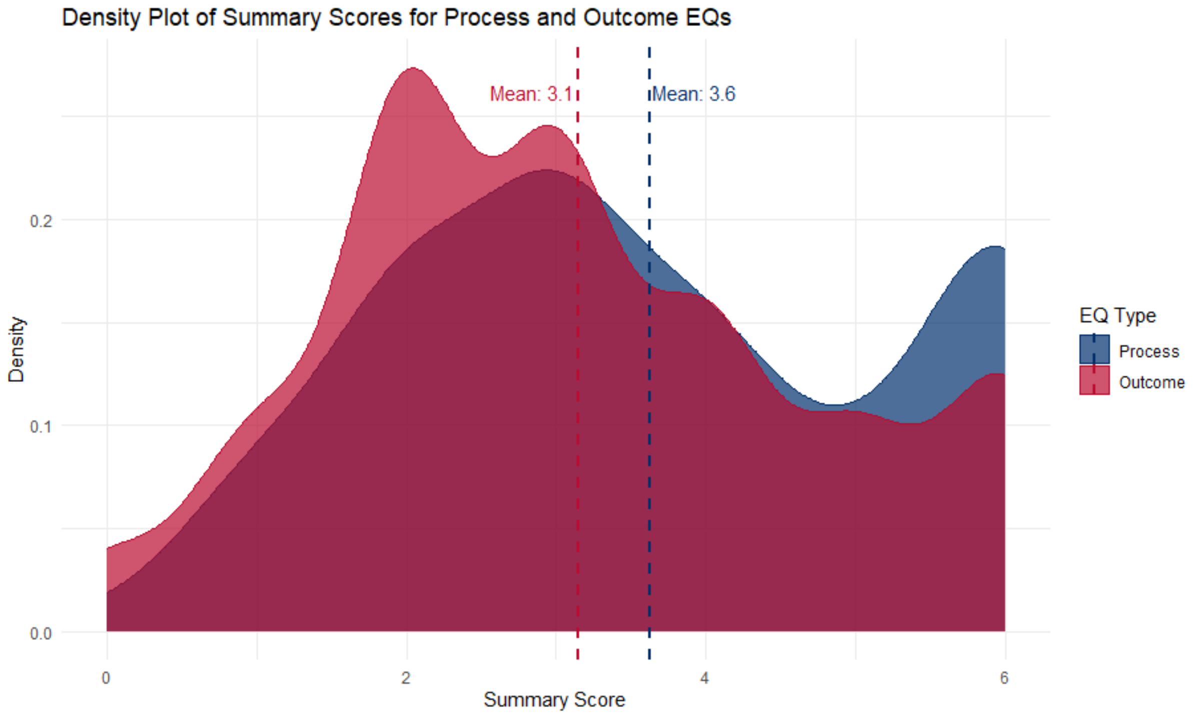
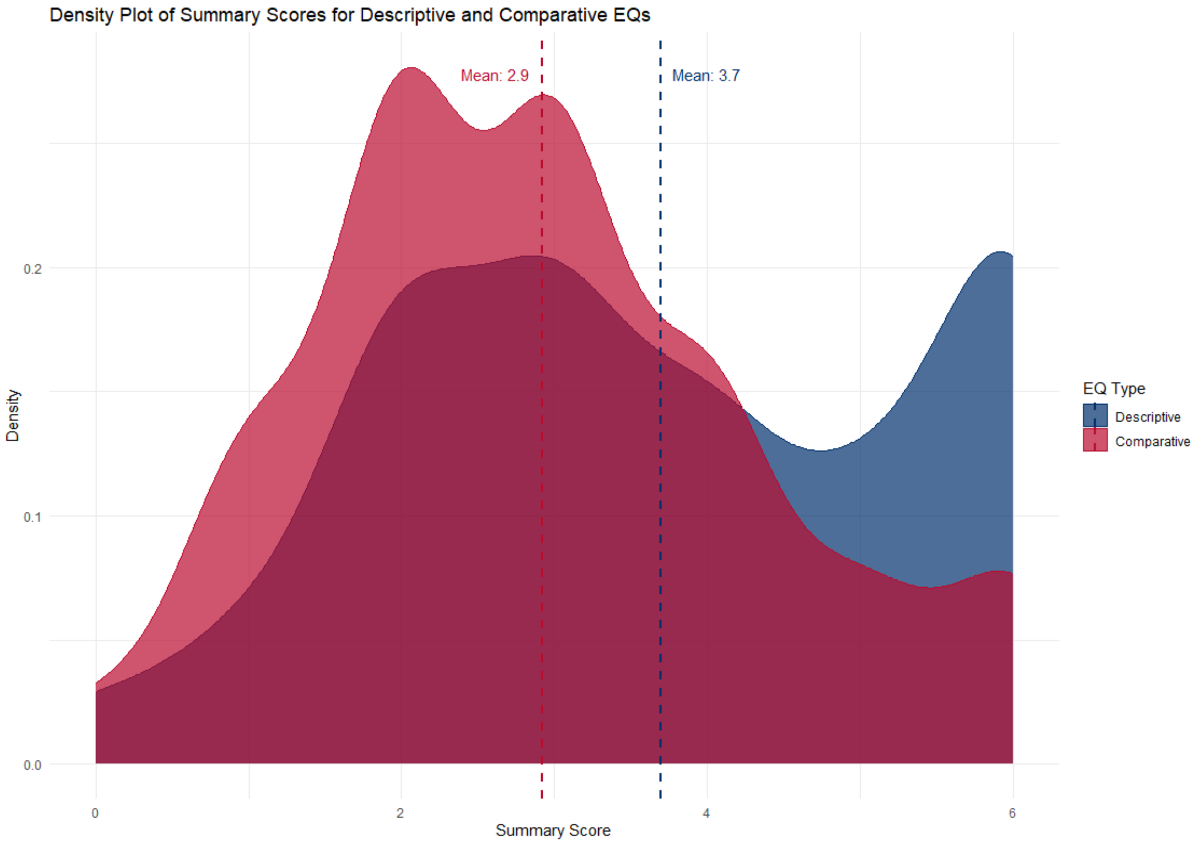


Figure 8: Density Plot Comparing Descriptive and Comparative Questions



APPENDIX 3: LIST OF LER PEs

The list of LER PE reports that the team reviewed during the study comes from the ALEC PE Inventory. When the report is available on the DEC, the team has included the link.

Table 2. LER PE Reports

ID	PE TITLE	MECHANISM	LEARNING PARTNER
1	Liberia—Governance and Economic Management Support Project Mid-Term (LER I NORC #5)	LER I	NORC
2	Cambodia—Countering Trafficking in Persons Mid-Term (LER I SI #9)	LER I	SI
3	Nepal—Peace Support Project Final Evaluation (LER I NORC #14)	LER I	NORC
4	Macedonia—Judicial Strengthening (LER I SI #7)	LER I	SI
5	Macedonia—Inter-Ethnic Education (LER I SI #11)	LER I	SI
6	Nepal—Peace Process and Constitutional Drafting Process Final Evaluation (LER I NORC #10)	LER I	NORC
7	Liberia—Land Conflict Resolution Project (LER I SI #10)	LER I	SI
8	Mozambique—Media Strengthening (LER I NORC #22)	LER I	NORC
9	Mozambique—Assistance to the Attorney General (LER I NORC #29)	LER I	NORC
10	Center-Run Elections and Political Processes Fund (LER I SI #13)	LER I	SI
11	Europe and Eurasia Bureau's Regional Investigative Journalism Network Program (LER I SI #14)	LER I	SI
12	Center-Run Consortium for Elections and Political Process Strengthening III LWA (LER I NORC #19)	LER I	NORC
13	Libya—DRG Program Mid-Term Evaluation (LER I SI #15)	LER I	SI
14	Ukraine—Media Strengthening Program (LER I SI #17)	LER I	SI
15	LGBTI Global Development Partnership (LER I NORC #65)	LER I	NORC
16a	Center-Run Displaced Children and Orphan's Fund (Moldova Programs) (LER I NORC #3)	LER I	NORC

ID	PE TITLE	MECHANISM	LEARNING PARTNER
16b	Center-Run Displaced Children and Orphan's Fund (Burundi Programs) (LER I NORC #3)	LER I	NORC
17	Syria PRIDE (LER I NORC #71)	LER I	NORC
18	Somalia Bringing Unity, Integrity, and Legitimacy to Democracy (LER I SI #23)	LER I	SI
19	Eastern and Southern Caribbean—Juvenile Justice (LER I SI #19)	LER I	SI
20	Kenya Yetu Community Philanthropy (LER I SI #25)	LER I	SI
21	Kenya SCORE Countering Violent Extremism (LER I SI #24)	LER I	SI
22	Sierra Leone WELD PE (LER II NORC #6)	LER II	NORC
23	Paraguay Governance Strengthening Final Performance Evaluation (LER I NORC #2)	LER I	NORC
24	Ukraine WOPE (LER I SI #26)	LER I	SI
25	Strengthening Civil Society Globally PE (LER II NORC #30)	LER II	NORC
26	Moldova WOPE (LER I SI #28)	LER I	SI
27	Morocco Civil Society Performance Evaluation (LER II NORC #25)	LER II	NORC
28	Malawi LGAP Mid-Term Performance Evaluation (LER II Cloudburst #20)	LER II	Cloudburst
29	Georgia GGI Mid-Term Performance Evaluation (LER II Cloudburst #25)	LER II	Cloudburst
30	Mali COVID-19 and Governance PE (LER I NORC #82)	LER I	NORC
31	Georgia Civil Society (ACCESS) (LER II NORC #40)	LER II	NORC
32	Comunitatea MEA Mid-Term Performance Evaluation (LER II NORC #55)	LER II	NORC
33	Côte d'Ivoire Political Transition and Inclusion Program Final Performance Evaluation	LER II	Cloudburst

ID	PE TITLE	MECHANISM	LEARNING PARTNER
34	Social Movements in Zimbabwe: A Field Assessment and Evaluation	LER II	Cloudburst
35	Zimbabwe Democracy and Governance Development Objective Final Performance Evaluation	LER II	Cloudburst
36	Ukraine DOPP Final Performance Evaluation	LER II	Cloudburst
37	Gender-Based Violence Portfolio Performance Evaluation: Collective Action to Reduce Gender-Based Violence Final Report	LER II	NORC
38	Gender-Based Violence Portfolio Performance Evaluation: Better Together Challenge Final Report	LER II	NORC
39	Performance Evaluation of the USAID Promoting Civic Education and Participation in South Africa Program	LER II	NORC
40	Gender-Based Violence Portfolio Performance Evaluation: RISE Final Report	LER II	NORC
41	Cambodia Social Accountability Portfolio Performance Evaluation Report	LER II	NORC
42	USAID/Peru TPI Integrity Networks Evaluation: Midline Evaluation Report	LER II	NORC
43	USAID/Guatemala Community Roots Activity: Final Performance Evaluation	LER II	NORC
44	Performance Evaluation of USAID's Response to COVID-19-Enabled Corruption: Final Report	LER II	NORC
45	Gender-Based Violence Portfolio Performance Evaluation: Women's Economic Empowerment Final Report	LER II	NORC
46	Mali Justice Project Performance Evaluation: Final Report	LER II	NORC
47	Nicaragua Civil Society and Political Processes Performance Evaluation	LER II	NORC
48	Center-Run Legally Enabling Environment Program (LER I NORC #7)	LER I	NORC
49	Burundi Youth, Conflict, and Peacebuilding (LER I NORC #44)	LER I	NORC
50	Liberia—Women's Leadership Study (LER I NORC #51)	LER I	NORC

ID	PE TITLE	MECHANISM	LEARNING PARTNER
51	Black Sea Trust (LER II NORC #4)	LER II	NORC
52	Ukraine Civil Society Mid-Term Evaluation (LER II NORC #5)	LER II	NORC
53	ISC Final Performance Evaluation (LER II NORC #12)	LER II	NORC
54	Nepal—Local Governance Program (LER I NORC #26)	LER I	NORC
55	Ukraine New Justice Mid-Term Performance Evaluation (LER II Cloudburst #14)	LER II	Cloudburst
56	Belarus Civil Society Mid-Term Performance Evaluation (LER II Cloudburst #16)	LER II	Cloudburst
57	Ukraine DG East Mid-Term Performance Evaluation (LER II Cloudburst #21)	LER II	Cloudburst
58	JRS HRSM RRM Mid-Term Performance Evaluation (LER II Cloudburst #24)	LER II	Cloudburst
59	Ukraine HRS Mid-Term Performance Evaluation (LER II Cloudburst #27)	LER II	Cloudburst
60	Ukraine MPU Mid-Term Performance Evaluation (LER II Cloudburst #32)	LER II	Cloudburst
61	Center-Run Global Labor Program Mid-Term Evaluation (LER I NORC #11)	LER I	NORC
62	Center-Run Information Safety and Capacity Project Mid-Term Performance Evaluation (LER I NORC #13)	LER I	NORC
63	Europe and Eurasia Strengthening Media Performance Evaluation (LER II NORC #22)	LER II	NORC

APPENDIX 4: INTERVIEWEES

The team identified 12 stakeholders for informational interviews. Of the original 12, nine agreed to participate.

Table 3. Stakeholders for Informational Interviews

STAKEHOLDER TYPE	NUMBER OF INTERVIEWS
USAID Washington (DRG, Bureau for Planning, Learning, and Resource Management)	4
USAID Mission staff who commission evaluations	3
Evaluation experts and MEL platform staff	2
Total interviews	9