PHOTO CREDIT: JEANNE HALLACY | SOLIDARITY CENTER

ADDRESSING LEARNING AND EVALUATION CHALLENGES (ALEC) REPORT

DEMOCRACY, HUMAN RIGHTS, AND GOVERNANCE LEARNING, EVALUATION, AND RESEARCH ACTIVITY III

AUGUST 2024

This publication was produced at the request of the United States Agency for International Development. It was prepared by Luis A. Camacho, Kate Marple-Cantrell, and Daniel Sabet.

ABSTRACT

Many evaluations and assessments involve a team of researchers conducting a series of key informant interviews, group interviews, and focus group discussions at one point in time over a three- to four-week period of fieldwork. Yet such studies regularly produce several complaints from commissioners, evaluators, and implementers related to the accuracy and reliability of the findings and the subsequent usefulness of the study. This study lays out guidance on addressing seven pain points in qualitative evaluation work, including: 1) case and site selection for small-n studies; 2) selection of respondents; 3) social desirability bias; 4) qualitative data capture; 5) qualitative data analysis; 6) evidentiary support for statements; and 7) clarity of findings to facilitate use. This guidance includes minimum standards to which all studies should adhere, good practices that studies should seek to implement when feasible, and guidance for study commissioners.

Tasking S3-03 / CB3-01
Contracted under GS-10F-0294V/7200AA23M00014 / GS-10F-0218U/7200AA23M00011

Submitted to:
Matthew Baker, USAID Contracting Officer's Representative (Social Impact)
Laura Berger, USAID Contracting Officer's Representative (Cloudburst)

Submitted by:
Mateusz Pucilowski, Chief of Party, Social Impact
Aleta Starosta, Chief of Party, Cloudburst

Contractors:
Social Impact
2300 Clarendon Blvd, Suite 1000, Arlington, VA 22201
Attention: Mateusz Pucilowski
Tel: 1-703-465-1884; E-mail: mpucilowski@socialimpact.com

The Cloudburst Group
8400 Corporate Dr #550, Landover, MD 20785
Attention: Aleta Starosta
Tel: +260 777380273; Email: aleta.starosta@cloudburstgroup.com

DISCLAIMER

CONTENTS

LIST OF TABLES AND FIGURES

ACRONYMS

ADS	Automated Directives System
AI	Artificial Intelligence
ALEC	Addressing Learning and Evaluation Challenges
AMELP	Activity Monitoring, Evaluation and Learning Plan
COP	Chief of Party
COR	Contracting Officer Representative
CSO	Civil Society Organization
DEC	Development Experience Clearinghouse
DRG	Democracy, Human Rights, and Governance
EQ	Evaluation Question
ET	Evaluation Team
FGD	Focus Group Discussion
IP	Implementing Partner
KII	Key Informant Interview
LER	Learning, Evaluation, and Research
LP	Learning Partner
MEL	Monitoring, Evaluation, and Learning
PE	Performance Evaluation
SI	Social Impact, Inc.
SDB	Social Desirability Bias
SOW	Scope of Work
USAID	United States Agency for International Development

EXECUTIVE SUMMARY

EVALUATION CHALLENGES

Many evaluations and assessments involve a team of researchers conducting several key informant interviews, group interviews, and focus group discussions at one point in time over a three- to four-week period of fieldwork. Yet such studies regularly produce several complaints from commissioners, evaluators, and implementers related to the accuracy and reliability of the findings and the subsequent usefulness of the study.

With learning partners (LPs) Social Impact and The Cloudburst Group, the United States Agency for International Development's Bureau for Democracy, Human Rights, and Governance (USAID/DRG) completed this study to develop guidance on addressing seven common pain points in qualitative evaluation work. These include: 1) case selection for small-n studies; 2) selection of respondents; 3) social desirability bias; 4) qualitative data capture; 5) qualitative data analysis; 6) evidentiary support for statements; and 7) clarity of findings to facilitate use.

This guidance includes: 1) minimum standards to which all studies should adhere and 2) good practices that studies should seek to implement when feasible. Through this inquiry, USAID/DRG also hopes to build consensus among USAID staff, LPs, and evaluators on these minimum standards and best practices. While these studies focus on performance evaluations in the DRG sector, guidelines should be widely applicable across sectors.

METHODOLOGY

The study team followed a sequenced approach, including: 1) identifying the team's initial thoughts about these issues (i.e., their priors) and building on the team's extensive experience; 2) reviewing recent DRG performance evaluation (PE) reports produced under the Democracy, Human Rights, and Governance Learning, Evaluation, and Research mechanism; 3) reviewing the academic and practitioner guidance literature; 4) conducting key informant interviews with evaluation and assessment commissioners, experts at key evaluator and research contracting firms, and past team leaders; and 5) convening workshops to review draft guidance and fill in the remaining gaps. After revising and finalizing the guidance, the team will conclude consensus-building efforts with a presentation of the guidance to USAID stakeholders.

EVALUATION PAIN POINTS: SUMMARY AND GUIDANCE

The below subsections summarize guidance for performance evaluations to address each pain point. Minimum guidance is denoted by this symbol ❗; other guidance contains suggested additional considerations and strategies. Additionally, a list of required elements for work plans and evaluation reports is included in Appendix 1: Required Elements.

1. IMPROVING SITE (CASE) SELECTION

The Challenge: Many evaluations, studies, and assessments typically entail selecting a small number of cases, units, or sites for deeper analysis. If not selected well, however, these can provide an inaccurate sense of the program or not adequately address questions and learning needs.

Guidance:

1.1 Decide whether the PE will require selecting sites as soon as possible after having the necessary information.

1.2 If selecting sites is necessary, determine the number of sites that the evaluation team can visit given the available resources and, to the extent possible, make this decision independently of actual site selection.

1.3 Determine the site selection approach and complete site selection only after developing a strong understanding of the activity and evaluation priorities.

1.4 Determine the site selection approach based on evaluation questions and the analytical goals of the evaluation. Consider two broad types of site selection approaches: representative and purposive.

1.5 Do not select sites using convenience sampling, but adjusting site selection to reflect security and accessibility considerations may be necessary.

1.6 Consult with USAID and implementing partner (IP) staff to inform the site selection strategy, but USAID and IP staff should not determine the actual site selection.

1.7 Consider whether the PE could benefit from an incremental or sequenced approach to site selection. If planning to use a sequenced approach, the justification and process to finalize site selection should be included in the work plan/design report.

1.8 Provide detailed information on sampling in the report, including any deviations from the original plans and their analytical implications.

2. IMPROVING RESPONDENT SELECTION FOR INDIVIDUAL AND GROUP INTERVIEWS

The Challenge: Most PEs and assessments rely on individual and group interviews, including those of program participants and indirect beneficiaries. PE reports often do not provide enough information about the selection strategy and how it aligns with EQs. This undermines trust in evaluation findings and creates a barrier to use.

Guidance:

2.1 Identify the potential population of respondents, including key informants, program participants, and indirect beneficiaries.

2.2 Identify potential key informants following an intentional process that draws on inputs from local team members, preliminary consultations with USAID and IP staff, and a thorough review of program documentation.

2.3 Leave room for flexibility in key informant selection.

2.4 Similar to site selection, select program participants and indirect beneficiaries using representative or purposive approaches as required by the evaluation questions.

2.5 To obtain breadth and depth, consider conducting a survey followed by a random or purposive selection of respondents for more in-depth interviewing.

2.6 Upon completing respondent selection, validate the selection against the evaluation questions and ensure that the information to be collected is sufficient to address them.

2.7 Provide detailed information on respondent selection in the report, including any deviations from original respondent selection plans and their analytical implications.

3. ADDRESSING SOCIAL DESIRABILITY BIAS

The Challenge: If an ET asks program participants if they are satisfied with a program or feel that it had a positive impact, many people will naturally answer "yes" regardless of their true perception. Despite this clear limitation, ETs frequently lean on these kinds of questions and risk drawing incorrect conclusions.

Guidance:

- 3.1 Recognize the difference between mitigation and resolution of social desirability bias (SDB).
- 3.2 Identify and use sources of data and data collection methods that are less subject to SDB.

3.3 Limit access to the data, including donor access, and communicate clearly to interviewees about how the data will be used.

- 3.4 In developing instruments, use indirect questioning, be thoughtful about question wording, and consider prefacing questions.

3.5 Pretest different approaches to questions subject to SDB.

- 3.6 During the interview itself, create a trusting atmosphere, look for cues of SDB, flag bias risks in notes, and follow up and probe.
- 3.7 Take SDB seriously in analyzing data, report writing, and quality assurance and caveat findings accordingly.

4. IMPROVING DATA CAPTURE AND MANAGEMENT

The Challenge: Interview and FGD notes are the data that findings and conclusions should be derived from, and yet in some cases, the data provided to research teams might not be well captured or stored to allow for meaningful analysis.

Guidance:

- 4.1 Develop and implement a data capture plan to consistently capture a verbatim or close-to-verbatim record of each qualitative event.

4.2 Find the right size in the mix of skills on the evaluation team.

- 4.3 Follow guidelines for informed consent and data protection, minimizing the collection of and access to raw or identifiable data.

4.4 Encourage teams to follow good practices for note-taking.

- 4.5 Absent extenuating political or security circumstances, practice the regular recording of qualitative events if consent is given and recording is unlikely to undermine frank responses.
- 4.6 Select LP staff should have access to all of their team's qualitative data and provide regular quality oversight.

5. IMPROVING DATA ANALYSIS

The Challenge: Without a systematic, documented, and somewhat replicable approach to data analysis, research teams risk several forms of bias, including confirmation bias and deriving conclusions from early interviews, recent interviews, or dynamic and memorable interviewees. In addition, it

becomes difficult for teams to collaborate, conduct quality control, or revisit findings and conclusions that are poorly documented.

Guidance:

5.1 Employ a documented systematic approach for arriving at findings from all data sources, including, at a minimum, structured thematic or content analysis for qualitative data.

5.2 Systematic analysis of qualitative data in PEs will most often be facilitated by first coding what was discussed into coherent categories.

5.3 Basic rapid analysis using manual thematic coding is sufficient in certain circumstances.

5.4 Some teams may want to pursue (and some Missions may request) coding using qualitative data analysis software.

5.5 LPs and team leads should provide leadership and oversight throughout the evaluation to ensure that the analysis plan is carried out faithfully.

5.6 Evaluations should be adequately budgeted and staffed to support systematic analysis.

5.7 AI is an emerging tool that, with caution, could be used to speed up the coding process or help uncover hidden themes.

5.8 Combine systematic analysis of qualitative data with other sources (e.g., desk review, surveys, activity monitoring, evaluation and learning plan data, and other secondary data) to triangulate findings.

6. IMPROVING THE EVIDENTIARY SUPPORT FOR STATEMENTS

The Challenge: There is conflicting guidance on the evidentiary support required for finding and conclusion statements. In some cases, tangential and poorly supported findings and conclusions are offered in reports, but proposed solutions to ensure evidentiary support often create other problems, such as requiring highly structured interviews, treating qualitative data like quantitative data, and ignoring many of the strengths of qualitative data.

Guidance:

6.1 Quantify qualitative data only when using highly structured instruments on a large sample.

6.2 The source of evidence should be cited in a way that provides basic information about the source while still maintaining confidentiality.

6.3 Findings must be based on multiple data points.

6.4 Reports should be organized by findings and not divided by data source.

6.5 Expectations about the extent and style of evidentiary support should be discussed early in the evaluation process and not wait until after a draft has been submitted.

6.6 Take additional steps to proactively build user confidence in the study findings and ensure their utility.

6.7 Be transparent in the level of confidence in findings.

7. IMPROVING THE PRESENTATION OF FINDINGS TO FACILITATE USE

The Challenge: Evaluations, studies, and assessment reports are often lengthy, and key points risk being buried in reports or never read by the intended users. Research teams often struggle to ensure that key points are highlighted without losing needed nuance or adequate empirical support.

Guidance:

7.1 Provide a summary of the question's response at the outset.

7.2 Use bolded topic sentences summarizing findings, followed by supporting evidence.

7.3 Use visualizations to summarize qualitative information.

7.4 Shift much of the methodology explanation to an annex.

7.5 Conclusions sections should identify implications for decision-making and where actions need to be taken.

7.6 Implement processes to ensure a well-written report, including a robust internal review process aided by a checklist.

7.7 Develop complementary products that go beyond the report, including targeted briefs, infographics, slides, presentations, videos, or podcasts.

BACKGROUND AND PURPOSE

Many evaluations, assessments, and studies commissioned under the Democracy, Human Rights, and Governance (DRG) Center's Learning, Evaluation, and Research (LER) mechanisms and other United States Agency for International Development (USAID) mechanisms involve a team of researchers conducting several key informant interviews (KIIs) and focus group discussions (FGDs) at one point in time over a two- to four-week period of fieldwork. These types of what might be called "traditional" performance evaluations (PEs) are the most common forms of external evaluation employed by USAID in the DRG sector. DRG assessments, political economy analyses, and similar qualitative studies also rely on these approaches. Yet, such studies regularly produce several complaints from commissioners, evaluators, and implementers related to the accuracy and reliability of the findings and the subsequent usefulness of the study.[1]

In recent years, advocates within the DRG learning community have argued for the use of more rigorous research methods, including impact evaluations and rigorous PEs that provide improved measures of outcomes, changes in outcomes over time, and attribution of programmatic impacts.[2] While the best way to improve DRG studies is to improve the rigor of the design, so-called traditional PEs that rely on KIIs, FGDs, and short stints of fieldwork are likely to remain popular.[3]

There have been several efforts in recent years to offer guidance, best practices, templates, and tools to improve evaluation, research, and assessment quality. Within USAID, two notable examples are the Learning Lab's Evaluation Toolkit and the Assessing the Quality of Education Evaluations tool from the USAID Office of Education. Outside of USAID, there are the Innovations for Poverty Action's Right-Fit Evidence unit, the World Bank's DIME Analytics Handbook, the BetterEvaluation platform, and the Office of Evaluation Statistics' Evaluation Resources, in addition to academic resources.[4]

The Addressing Learning and Evaluation Challenges (ALEC) study does not intend to duplicate these efforts or provide start-to-finish guidelines for conducting traditional PEs and other qualitative assessments and studies, hereafter collectively referred to as "traditional PEs." Instead, this study focuses on problem areas or "pain points" that commonly arise and limit the credibility of findings from traditional PEs. By focusing on these areas, this study takes a deeper dive into how to address these pain points within existing budgetary and time constraints. In addition to developing guidance on minimum standards and best practices, the study aims to build consensus among USAID staff, implementing partners (IPs), and researchers on these minimum standards and best practices.

The intended users of the study are USAID staff commissioning, managing, and using traditional PEs and IPs and researchers conducting these evaluations. While the study focuses on traditional PEs in the DRG sector, guidelines are widely applicable across sectors.

[1] Accuracy refers to the correctness and unbiasedness of the findings (i.e., are an evaluation's findings "true"?) while reliability refers to the trustworthiness and consistency of the findings (i.e., are an evaluation's findings supported by evidence, and would another similar evaluation reach the same findings?).

[2] Findley, M. G., Starosta, A., & Sabet, D. (2022). *DRG impact evaluation retrospective: Learning from three generations of impact evaluations*. USAID. https://pdf.usaid.gov/pdf_docs/PA00XF3F.pdf

[3] Reasons for the popularity of traditional PEs include that they are flexible and do not require evaluation planning at the activity design stage, USAID staff and external evaluators have considerable experience with this type of design, such evaluations are very good at answering certain types of questions (e.g., lessons learned), they are lower cost, they meet the ADS requirements, and their findings tend to be more positive than the findings of IEs.

[4] For an example on case selection, see: Seawright, J., & Gerring, J. (2008). Case selection techniques in case study research: A menu of qualitative and quantitative options. *Political Research Quarterly, 61*(2), 294-308.

OBJECTIVES AND RESEARCH QUESTIONS

The objectives of this study are twofold:

- Develop guidance for LER III partners and other researchers to address recurring pain points that arise when conducting traditional PEs. This includes: 1) minimum guidance to which all studies should adhere—denoted by this symbol ❗ in the sections below and 2) additional good practices that studies should seek to implement when feasible.
- Build consensus among key individuals and organizations involved in such studies on the minimum guidance and best practices.

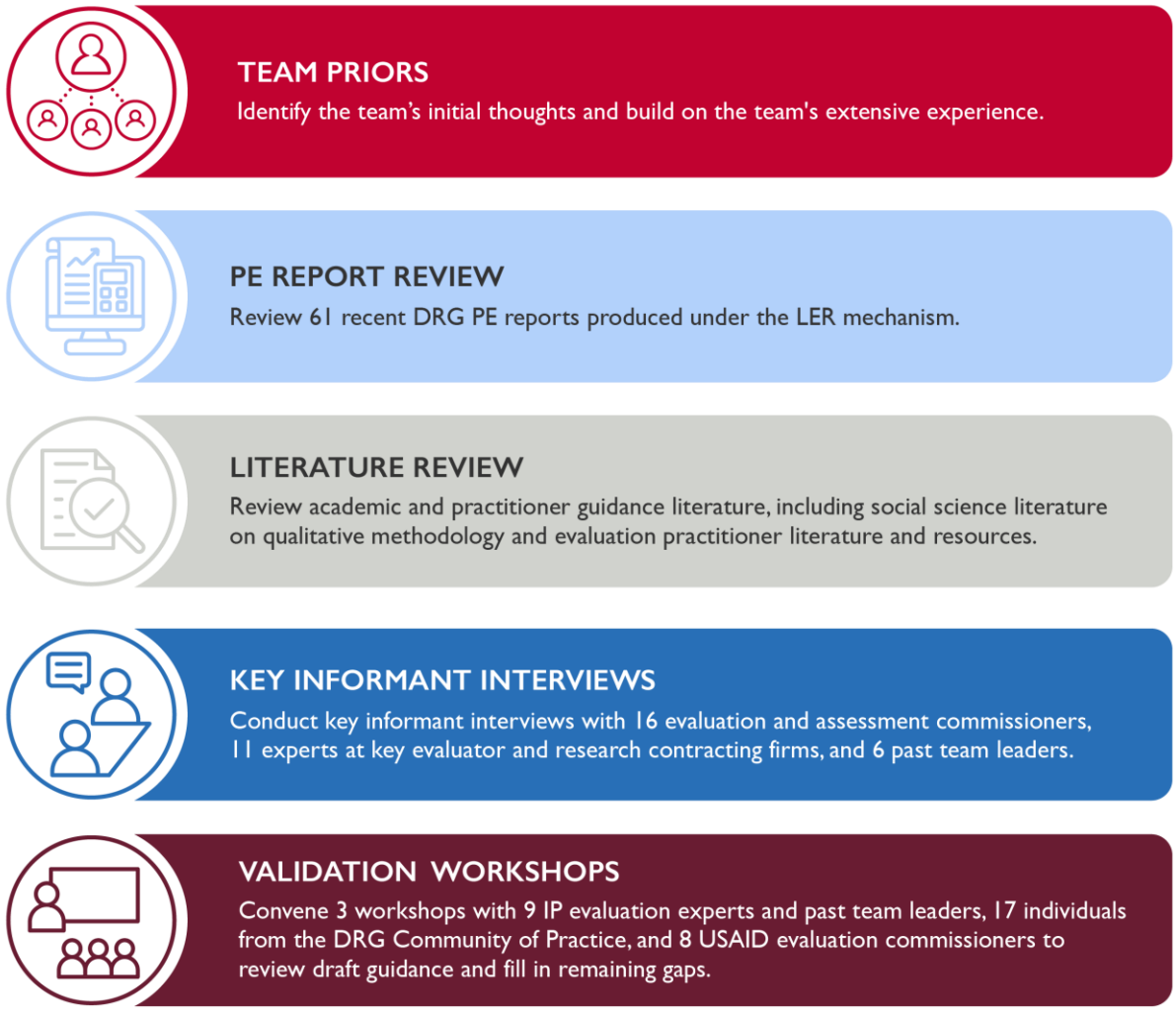
In lieu of research questions, the study focuses on the pain points determined by the DRG Bureau's Evidence and Learning team, listed below:

1. Case or site selection for small-n studies
2. Selection of respondents
3. Social desirability bias (SDB)
4. Qualitative data capture
5. Qualitative data analysis
6. Evidentiary support for statements
7. Clarity of findings to facilitate use

APPROACH

As seen in Figure 1, the study team[5] followed a sequenced approach, including: 1) identifying the team's initial thoughts about these issues (i.e., their priors) and building on the team's extensive experience; 2) reviewing recent DRG PE reports produced under the LER mechanism; 3) reviewing the academic and practitioner guidance literature; 4) conducting KIIs with evaluation and assessment commissioners, experts at key evaluator and research contracting firms, and past team leaders; and 5) convening workshops to review draft guidance and fill in the remaining gaps. After revising and finalizing the guidance, the team will conclude consensus-building efforts with a presentation of the guidance to USAID stakeholders. For more information, please refer to Appendix 2: Full Description of Study Methods.

Figure 1: Approach



[5] The study team comprised Luis A. Camacho, Senior Technical Director, Social Impact, Inc.; Kate Marple-Cantrell, Senior Evaluation Specialist, The Cloudburst Group; and Daniel Sabet, Senior Learning Advisor, USAID. Carolyn Lynch, Senior Research Assistant, Social Impact, Inc.; Kaila Cook, Senior Program Associate, Social Impact, Inc.; and Kellie Haddon, Junior Research Analyst, The Cloudburst Group provided management and research assistance.

FINDINGS AND RECOMMENDATIONS

GENERAL GUIDANCE

Before presenting findings and recommendations pertaining to the seven pain points, this section highlights some general, cross-cutting guidance relevant to all PEs that can help address several specific pain points. This guidance reflects inputs gathered from documents, interviews, and feedback on earlier versions of this guide.

EVALUATION QUESTIONS

Many of the seven pain points the ALEC study focuses on are exacerbated by scopes of work (SOWs) that include too many evaluation questions (EQs) or questions that are too broad in scope. The SOW development and peer review process should ensure that questions are tied to use needs, narrow in scope, clear, and feasible to answer. USAID's Learning Lab offers some useful resources for developing EQs, and another ALEC study is developing a forthcoming workbook to aid in developing questions. LPs and evaluation teams (ETs) should discuss with USAID any issues with the EQs as soon as possible after receiving an SOW to ensure that the final set of questions can be addressed given available resources.

STAFFING AND LOCAL KNOWLEDGE

USAID and partners responsible for PEs should ensure that ETs have the necessary expertise to develop and implement an evaluation design that is fit for purpose. This entails that the evaluation: 1) addresses EQs; 2) meets USAID evaluation requirements, the guidance in this document, and other relevant standards; and 3) is grounded on knowledge of the local context. This includes the evaluation and research design expertise of evaluation specialists and social scientists, as well as the knowledge and skills brought by subject-matter experts and local team members. As stated in Automated Directives System (ADS) chapter 201.3.6.8:

- Evaluations must be conducted by individuals with appropriate training and experience, including, but not limited to, evaluation specialists and technical or subject matter experts, including local experts. To the extent possible, evaluation specialists with appropriate expertise from the partner country, but who are not involved in project implementation, should lead and/or be included in the evaluation team.

It is key that team leads have evaluation expertise in addition to subject-matter expertise. Team leads will be responsible for designing evaluations, including carrying out site and interview selection, developing data capture and analysis protocols, overseeing their implementation in the field, and producing final deliverables. In addition to selecting the right team leads, it is also key to ensure the meaningful participation of local evaluation or subject-matter experts in all evaluation stages.

TRAINING AND OVERSEEING ETs

While an adequate staffing structure and recruitment efforts are important elements to ensure the quality of PEs, LPs should also plan to train ETs as needed. LPs should diagnose the training needs of ET members and ensure that they are equipped with the necessary knowledge and skills to implement high-quality PEs, including the minimum guidance discussed in this guide. Key areas to ensure training include identifying and responding to potential SDB during interviews, analyses, and writings; taking proper interview notes; organizing notes and other evidence; developing coding templates or codebooks; and using findings, conclusions, and recommendations matrices.

LPs must also provide technical and management oversight throughout the evaluation period of performance. This includes ensuring that ETs implement evaluation design and analysis plans faithfully, follow security protocols, and upload interview notes to servers regularly; reviewing all deliverables

prior to submission; and checking compliance with the minimal standards in this guidance. LPs should also be prepared to provide research support to teams as needed.

This recommended level of capacity building and oversight will require evaluation budgets that can accommodate LP staff in addition to the ET.

1. SITE (CASE) SELECTION

SUMMARY

The Challenge: Many evaluations, studies, and assessments typically entail selecting a small number of cases, units, or sites for deeper analysis. If not selected well, however, these can provide an inaccurate sense of the program or not adequately address questions and learning needs.

Minimum Guidance:

- Determine the site selection approach based on EQs and the analytical goals of the evaluation (1.4).
- Consider two broad types of site selection approaches: representative and purposive (1.4).
    - Representative site selection is intended to produce a subset of sites that resembles the characteristics of the population of sites.
    - Purposive site selection is intentional and considers site characteristics, selecting specific sites to best achieve analytical goals.
- Do not select sites using convenience sampling, but adjusting site selection to reflect security and accessibility considerations may be necessary (1.5).
- Consult with USAID and IP staff to inform the site selection strategy, but USAID and IP staff should not determine actual site selection (1.6).
- Describe the site selection approach in the main body of the evaluation report and include detailed information in an annex. Note any deviations from the original plans and the analytical implications of these deviations (1.8).

THE CHALLENGE

DRG programming often targets multiple distinct geographies or (sub)sectors within countries. For example, local governance activities usually work with a subset of subnational units, and state capacity-building activities usually work with a subset of government agencies. Other DRG programming, often managed centrally by the DRG Bureau, is implemented across countries or regions. Yet, another type of programming consists of an interrelated set of activities intended to achieve similar objectives within a country or across several countries. Depending on the type of programming, these subnational geographies, subsectors, countries, regions, and activities can be thought of as distinct cases, units, or sites (hereafter referred to as "sites").

PEs of DRG programming implemented across more than a handful of sites often must select a subset of sites to focus on during field research. When discussing site selection in interviews, the study team found general agreement that PEs should clearly describe their site selection approach and that selection should be grounded in the EQs. However, the review of PE reports found that reports often fail to meet these expectations. The study team found that just over half of the evaluations that selected sites for field research provided a justification for the selection approach (24 of 46 evaluations). Among those that provided a justification, the vast majority (19 of 24) indicated that their selection was purposive, but the level of detail provided varied substantially and was often not enough to fully understand the selection approach. Three evaluations selected sites based on convenience, which is generally considered to be of limited analytical value and was mentioned as a concern in three interviews.

Lastly, ETs seeking to do site selection "right" face several challenges. Site selection often takes place early in the evaluation and sometimes before ETs have had a chance to develop a thorough understanding of the activity and obtain a list of all sites (i.e., a list of the population of sites). Limited understanding of activities and incomplete information can lead to ETs being over reliant on IPs and

USAID recommendations for site selection. Moreover, PEs generally include multiple and complex EQs with corresponding sub-questions that might require different site selection strategies.

GUIDANCE FOR ETs AND LPs

1.1 Decide whether the PE will require selecting sites as soon as possible after having the necessary information. This decision should be taken after receiving a complete list of sites from the IP or better yet, a database containing site characteristics, site-disaggregated implementation, and performance monitoring information. Ideally, USAID and the IP would prepare such content prior to issuing an evaluation SOW. If a list or database is not available, the ET should seek to obtain information about sites as soon as possible. The decision should consider the PE timeline and available resources, as well as client priorities. In addition to SOWs, kick-off calls and initial exploratory interviews with USAID and IP staff could be useful to inform this decision. If there is a small number of sites (less than six to eight is a good rule of thumb), consider going to all sites and avoid site selection issues altogether.

1.2 If selecting sites is necessary, determine the number of sites to visit given the available resources and, to the extent possible, make this decision independently of actual site selection. This selection could be used for planning and budgeting purposes early in the evaluation design and relieve the pressure on finalizing the actual case selection. Consider visiting as many sites as possible while keeping in mind that the value added from additional sites might decrease after a certain point. Indeed, too much data from too many sites could make analysis unmanageable, and, as mentioned above, it might be more analytically valuable to develop a deeper understanding of a few sites. Importantly, even if most sites are selected, site heterogeneity (or heterogenous implementation across sites) might make it difficult to generalize findings to all sites.

1.3 Determine the site selection approach and complete site selection only after developing a strong understanding of the activity and evaluation priorities. This includes receiving and reviewing documentation and performance monitoring data and conducting preliminary interviews and discussions with USAID and IP staff. Operationally, this will likely entail developing the site selection strategy at the end of the work plan/design development stage.

1.4 Determine the site selection approach based on EQs and the analytical goals of the evaluation. Consider two broad types of site selection approaches: representative and purposive. Note that these two broad types are one of many possible selection approach typologies.[6] Also, note that a given PE might rely on a combination of approaches.

Representative Site Selection: Selection is intended to produce a subset of sites that resembles the characteristics of the population of sites. There are at least two ways to operationalize this approach:

1. *Stratified (random) selection:* This entails dividing the population of sites into smaller groups (i.e., strata) of relatively similar sites using one or more characteristics and randomly choosing sites

[6] Other possible typologies include theory- or hypothesis-driven approaches that ground site selection on analytical goals (Patton, M. Q. (2002). *Qualitative evaluation checklist*. Evaluation checklist project. The Evaluation Center, Western Michigan University; Goertz, G., & Mahoney, J. (2012). *A tale of two cultures: Qualitative and quantitative research in the social sciences*. Princeton University Press; Humphreys, M., & Jacobs, A. M. (2023). *Integrating inferences: Causal models for qualitative and mixed-method research*. Cambridge University Press; Seawright, J., & Gerring, J. (2008). Case selection techniques in case study research: A menu of qualitative and quantitative options. *Political Research Quarterly, 61*(2), 294-308; Collier, D., Brady, H. E., & Seawright, J. (2011). A sea change in political methodology. *Newsletter of the American Political Science Association, 9*(1).) and practical or "realistic" approaches that ground site selection on the feasibility of implementing specific procedures (e.g., US General Accounting Office. (1990). *Case study evaluations*. https://www.betterevaluation.org/sites/default/files/10_1_9.pdf; USAID. (2013b). *Evaluative case studies.* [Technical note]. Version 1.0. Monitoring and Evaluation Series.). The discussion below does not include "most similar" and "most different" systems design for comparative case studies (Seawright, J., & Gerring, J. (2008). Case selection techniques in case study research: A menu of qualitative and quantitative options. *Political Research Quarterly, 61*(2), 294-308; Collier, D., Brady, H. E., & Seawright, J. (2011). A sea change in political methodology. *Newsletter of the American Political Science Association, 9*(1).). Responsible use of these approaches requires developing deep knowledge of cases prior to selection, which is likely not feasible within the scope of typical PEs.

within each group if the group is larger than the number of sites that can be visited.[7] Alternatively, to random selection, ETs could use another criterion—such as the typical site described below—to finalize selection within each stratum. For example, a USAID activity might be implemented in 21 sites distributed across three regions and follow a phased approach whereby some sites receive four to five years of programming and other sites receive two to three years. Stratification for such a program could be depicted in Table 1. After dividing the sites into six strata, and having determined that they can visit six sites, the ET would select Site 8 in the Region B–Phase One stratum and randomly choose one site in each of the other five strata.[8]

Table 1: Illustrative stratification of 21 activity sites

PHASES	REGION A	REGION B	REGION C
One (four to five years of support)	Site 1 Site 2 Site 3	Site 8	Site 15 Site 16 Site 17 Site 18
Two (two to three years of support)	Site 4 Site 5 Site 6 Site 7	Site 9 Site 10 Site 11 Site 12 Site 13 Site 14	Site 19 Site 20 Site 21

Stratified selection works well when there are two characteristics or criteria to classify sites, and it is clear what those two criteria should be. While it is possible to consider additional criteria, adding a single criterion will double the number of strata and selected sites. Importantly, selecting different criteria will result in different site selections, so ETs should justify their criteria selection in addition to describing them.

2. *Typical site selection:* Sites are selected because the ET judges them to be representative of the "typical" or "average" site in the population of sites.[9] For example, after reviewing the available documentation and conducting exploratory interviews with IP staff and USAID, the ET determines that sites can be classified into three groups: a small group of sites where activity implementation stopped due to unforeseen circumstances, a larger group of sites where implementation proceeded as planned, and another small group of sites where implementation was particularly challenging. Wishing to learn about the typical implementation experience, the ET chooses to conduct field research in sites in the second group.

Importantly, as explained in Box 1.1, purely random site selection will likely not produce a representative sample of sites for the typical PE.

[7] Stratified (random) selection is similar to "diverse" selection (Seawright, J., & Gerring, J. (2008). Case selection techniques in case study research: A menu of qualitative and quantitative options. *Political Research Quarterly, 61*(2), 294-308) and "typology" selection (USAID. (2013b). *Evaluative case studies.* [Technical note]. Version 1.0. Monitoring and Evaluation Series.) approaches, whereby sites are selected to represent the full range of combinations of characteristics in the population of sites.

[8] For an application of this approach in an actual PE, see the USAID/Malawi Local Government Accountability and Performance Midterm Performance Evaluation Report from The Cloudburst Group. The ET chose 10 of Malawi's 28 districts for the Midterm Performance Evaluation of USAID's Local Governance Accountability and Performance (LGAP). Districts were stratified and then randomly selected within the strata. The ET used the following criteria for stratification: 1) wave of LGAP implementation (first or later); 2) region (north, center, and south); and 3) performance of local government (as measured in a pre-activity assessment).

[9] Seawright, J., & Gerring, J. (2008). Case selection techniques in case study research: A menu of qualitative and quantitative options. *Political Research Quarterly, 61*(2), 294-308.

Box 1.1: Why pure random site selection is not a good idea for typical Pes

> Pure random sampling results in representative samples when the population is large, and the sample to be selected is of sufficient size relative to the population. Hardly any PEs will meet these requirements. If the population and the number of sites to be selected are small, pure random selection is very likely to produce samples of sites that are not representative of the population.[10]

Purposive Site Selection: Like representative site selection, purposive selection is intentional and considers site characteristics. However, purposive selection does not seek to produce a subset of sites that resembles the full set of sites. It selects specific sites to best achieve analytical goals. There are many ways to do purposive site selection. Three selection examples are:

1. Extreme site selection (or bracketing): Sites are selected because they exemplify an extreme or unusual characteristic or combination of characteristics.[11] One extreme site selection strategy that could be useful for PEs is to rank sites according to performance or results metrics and choose the top and worst performers to learn about success drivers.

2. Intensity-based site selection: Sites are selected because they display a characteristic or set of characteristics of particular interest in an intense (but not extreme) way.[12] One intensity-based site selection strategy that could be useful for PEs is to select an activity's priority sites—e.g., a set of locations or government agencies that received programming in support of all the activity's results and sub-results. Another strategy could be to rank sites according to performance or results metrics and choose those that performed above average while excluding top-performing outliers.

3. Stratified purposive selection: This entails dividing the population of sites into smaller groups, as described above. However, stratified purposive selection selects sites from strata that are of particular interest given EQs rather than selecting sites in all strata in a representative way. ETs could use this strategy to ensure they reach sites with vulnerable populations (e.g., ethnic minority areas, low-income peri-urban areas, etc.).

The bullets below offer some general guidance on selection strategies tailored to evaluation goals:

- If the main goal of the evaluation is to learn from specific types of sites, and there is sufficient information to guide selection, consider a purposive strategy.
- If the main goal of the evaluation is to learn about how the program is being implemented and experienced by the target population, use a representative strategy.
- If the main goal of the evaluation is unclear, or if site selection needs to balance competing goals that cannot be prioritized, use a representative strategy that is both geographically and "programmatically" representative. Being programmatically representative entails selecting sites from across the variations or flavors created by uneven (whether intended or unintended) activity implementation.[13]

[10] Seawright, J., & Gerring, J. (2008). Case selection techniques in case study research: A menu of qualitative and quantitative options. *Political Research Quarterly, 61*(2), 294-308.

[11] Seawright, J., & Gerring, J. (2008). Case selection techniques in case study research: A menu of qualitative and quantitative options. *Political Research Quarterly, 61*(2), 294-308; USAID. (2013b). *Evaluative case studies.* [Technical note]. Version 1.0. Monitoring and Evaluation Series.

[12] Patton, M. Q. (2002). *Qualitative evaluation checklist*. Evaluation checklist project. The Evaluation Center, Western Michigan University.

[13] If implementation is standardized across sites (i.e., if implementation is homogenous), the programmatic representation criterion is not relevant.

1.5 Do not select sites using convenience sampling, but adjusting site selection to reflect security and accessibility considerations may be necessary. Ideally, original site selection would include additional "replacement" sites to be used if sampled sites are not accessible. This replacement strategy would preserve the integrity of the original selection strategy. In some cases, adjustments might need to happen before site selection—e.g., when a large collection of sites or a whole region is inaccessible due to conflict. This would effectively redefine the population of sites for the PE and should be noted in the report along with any potential implications on the generalizability of the findings.

1.6 Consult with USAID and IP staff to inform the site selection strategy, but USAID and IP staff should not determine actual site selection. As mentioned above, PEs often include EQs that require different selection approaches, so site selection might require prioritizing one question over others or developing a selection approach that can address more than two questions. In other PEs, ETs might need to build consensus around site selection criteria, or information about site characteristics might be limited, so consultations might provide an opportunity for the ETs to understand sites. While considering USAID's and IP's inputs is important, ETs should not let them drive actual site selection as this could introduce biases—e.g., both parties might want to direct the ET to sites where the activity has shown positive results or to sites that they are most familiar with.

1.7 Consider whether the PE could benefit from an incremental or sequenced approach to site selection. If planning to use a sequenced approach, the justification and process to finalize site selection should be included in the work plan/design report. In PEs where site selection needs to take place with incomplete information, an incremental approach might be a necessity. For example, the first phase of the evaluation would be used to inform case selection for a subsequent fieldwork phase. In other PEs, the flexibility of an incremental approach might be desirable for analytical purposes. Indeed, data collected in the early stages of the PE can suggest new lines of inquiry and corresponding sites for research.[14]

For example, during initial site visits, an ET might observe that sites where the activity under evaluation partnered with grass-roots organizations achieved the best results. Wanting to further explore this working hypothesis, the ET might change the site selection strategy to include the presence and absence of grass-roots organizations as a site selection criterion. Another variation of the incremental approach relies on mixed methods, whereby an initial survey in all sites informs subsequent site selection for field research.

1.8 Provide detailed information on sampling in the report, including any deviations from the original plans and their analytical implications. The methods section in the main body of the evaluation report should provide a general overview of site selection, including information about the population, selection approach, and selection criteria. Deviations from original site selection plans and the analytical implications of these deviations should also be noted in the methods section. The methods annex should include detailed site selection information. The information provided in the annex should be sufficient to allow for the replication of the selection procedure in combination with the database used for selection, which does not need to be included in the annex.

[14] One label for this approach is "opportunistic" or "emergent" sampling (Patton, M. Q. (2002). *Qualitative evaluation checklist*. Evaluation checklist project. The Evaluation Center, Western Michigan University.).

GUIDANCE FOR STUDY COMMISSIONERS

1. Doing site selection "right" requires time for the ET to develop familiarity with the activity and sites. This requires commissioning PEs six to eight months before a final report is needed and allowing four weeks between the kickoff call and the evaluation work plan submission. If SOW preparation becomes a bottleneck, consider issuing a less-detailed SOW, including an evaluation design phase. During this phase, the ET would work with the Mission (and IP) to develop a design that would become the SOW for the evaluation implementation phase. This design phase should include codesign meetings (or, better yet, a workshop, assuming sufficient funding is available) as well as a presentation of the evaluation work plan/design.

2. Prepare for the PE by obtaining needed information for site selection from the IP to share with the ET together with other project documentation (e.g., annual and quarterly reports). The minimum required information should be a list of all sites, including basic characteristics (e.g., the region) and key information about implementation (e.g., phase or intensity of implementation). Consider requiring IPs to collect and report performance monitoring data disaggregated by site and sharing this information with the ET to inform site selection.

2. RESPONDENT SELECTION FOR INDIVIDUAL AND GROUP INTERVIEWS

SUMMARY

The Challenge: Most PEs and assessments rely on individual and group interviews, including those of program participants and indirect beneficiaries. PE reports often do not provide enough information about the selection strategy and how it aligns with EQs. This undermines trust in evaluation findings and creates a barrier to use.

Minimum Guidance:

- ETs should map all types of respondents, including key informants, program participants, and indirect beneficiaries (2.1).
- ETs should identify potential key informants following an intentional process that draws on inputs from local team members, preliminary consultations with USAID and IP staff, and a thorough review of program documentation (2.2).
- Similar to site selection, ETs should select program participants and indirect beneficiaries using representative or purposive approaches, in line with EQs. Regardless of the selection approach, ETs should follow a process that aims to capture the full range of perspectives around an issue among the target respondent group (2.4).
- ETs should keep track of nonresponses and be attentive to its potential implications on their findings (2.7).
- ETs should describe the respondent selection approach in the main body of the evaluation report and include detailed information in an annex (2.7).

THE CHALLENGE

PEs do not always provide clear information about how they selected respondents for individual and group interviews. Just over half of the reviewed PEs (31 of 59) provided an explanation of how they selected respondents other than those that were interviewed because of their role—e.g., chiefs of party (COPs) and contracting officer representatives (CORs). This concern was also mentioned in three interviews. The lack of clear information about the respondent selection approach might leave evaluation users wondering if the selection process could have systematically excluded important groups, undermining the accuracy of findings.

PEs usually do not differentiate between proper key informants and other respondents, especially program participants and indirect beneficiaries (see Box 2.1 for a definition), when describing their selection strategies.[15] Key informants are specific individuals who have direct, first-hand knowledge about a topic of interest for the PEs. Other respondents are those who belong to a certain group (i.e., population) of interest, like activity or training participants, grantees, trainers, and technical assistance providers. Within each of these groups, individuals have unique perspectives or insights to share, but in principle, no single individual has more valuable information than others. For simplicity, the report refers to these other respondents as "program participants" and "indirect beneficiaries" henceforth.

[15] While the PE review did not quantify how many reports distinguished between key and other informants, the study team's impression is that the bulk of PEs use the term KII to refer to interviews that are 60 to 90 minutes long and often semi-structured, regardless of who the intended respondents are.

Box 2.1: Defining participants and indirect beneficiaries

> Program participants are individuals, collections of individuals, or organizations that are directly engaged with and benefit from an activity. Typical examples of DRG programming participants are civil society organizations (CSOs), media outlets and journalists, and government agencies and staffers. Indirect beneficiaries are those who benefit indirectly from the goods and services provided to or created with program participants. Indirect beneficiaries vary from activity to activity and include those who engage with CSOs, the media, government agencies, participants of other DRG programming, and the public. For example, the program participants in a local governance activity might be the local government or CSO staff members who participate in training or mentorship programs, while the indirect beneficiaries might be the citizens.

Key informant selection usually involves an identification or mapping process, not a selection of a few informants from a larger population. By contrast, the selection of program participants and indirect beneficiaries should be done in a transparent and unbiased way. A particular challenge for selecting program participants and indirect beneficiaries is that ETs are often unable to access complete lists. The unavailability of complete lists often compounds the difficulties in contacting and mobilizing them, especially when they are members of marginalized groups—e.g., at-risk youths, members of the LGBTIQ community, and survivors of gender-based violence.

This situation can result in ETs having to rely on IPs to mobilize program participants or on indirect beneficiaries for interviews or FGDs. If USAID and especially IPs are selecting these respondents, the ET might end up interviewing individuals who have more positive views about the program than the average participant or indirect beneficiary. The ET might also end up hearing only from those program participants who are the most engaged.

GUIDANCE FOR ETs AND LPs

2.1 Identify the potential population of respondents, including key informants, program participants, and indirect beneficiaries. Key informants can be "internal" and "external" to the activity. Internal key informants include USAID and IP staff, key government partners, and key program participants, if any—e.g., the representative of a CSO that participated in multiple activity interventions. External key informants are those who are not directly involved in the activity but are knowledgeable enough to provide a perspective, such as academics and other experts, government officials, journalists, and other donors. The perspectives of these external informants or activity outsiders can enrich evaluation findings. Other informants include program participants and indirect beneficiaries, and this will vary across activities.

Note that the key informant and other informant categories are not necessarily mutually exclusive. For example, in some instances, specific program participants might be key informants because they have taken part in multiple programming or have been engaged in an activity for a long time. In such cases, these participants would be included as key informants and selected through a purposive approach as described in 2.2.

2.2 Identify potential key informants following an intentional process that draws on inputs from local team members (see Box 2.2), preliminary consultations with USAID and IP staff, and a thorough review of program documentation. As ETs are aware, some ex officio key informants like COPs, deputy COPs, and CORs are predetermined for most PEs. To identify other key informants, ETs should cast as wide a net as possible to avoid excluding important internal and external perspectives.

Box 2.2: Local knowledge for respondent identification and inclusion

Local knowledge is key to selecting respondents, especially key informants that are external to an activity and might bring outside, fresh perspectives. Local knowledge is also useful for understanding the power dynamics that might make some respondents unwilling to participate or share their views. For example, in some contexts, a person of high authority, like a district commissioner, may not be a very useful respondent, but people under them could be. Equipped with local knowledge, the ET could decide to interview the authority for protocol reasons, to avoid upsetting them, and to ensure access to the actual intended respondents under them. ETs should also be attentive to what might be needed to ensure the inclusion of program participants, especially gender and resource considerations. For example, in some cases, ETs might need to take specific steps to include women (e.g., providing childcare).

If initial identification results in more potential key informants that can be interviewed given the available resources, ETs should prioritize those informants that offer to provide information that is most relevant to answer the EQs. If prioritization is necessary, it should be disclosed when describing respondent selection in the evaluation report. If the exclusion of some lower priority respondents could have substantial implications on findings, this should be disclosed in the limitations section of the report, as well as when presenting those findings that are impacted by the missing interviews.

2.3 Leave room for flexibility in key informant selection. This recognizes that ETs will increase their knowledge of the activity as they conduct the evaluation and become increasingly capable of identifying who might be the most knowledgeable informants and adjusting their selection accordingly.[16] Snowballing, whereby ETs identify additional interviewees with the assistance of interviewed key informants, is one way to keep the selection flexible. This can be particularly helpful in identifying external key informants. Another strategy is to do the selection in an iterative way, whereby ETs pause and reflect on findings to date, determine new or deeper lines of inquiry, and identify appropriate informants to pursue them. If flexibility in key informant selection is desired, ETs should plan for an additional two to three weeks for data collection.

2.4 Similar to site selection, select program participants and indirect beneficiaries using representative or purposive approaches as required by the EQs. Such a process would entail three main steps: 1) developing a list of all group members—i.e., a sampling frame; 2) stratifying the population; and 3) selecting potential respondents.

1. *Sampling frame development:* Ideally, IPs would keep complete lists of program participants (and indirect beneficiaries) as part of their monitoring, evaluation, and learning (MEL) database and share them with ETs at the beginning of each evaluation. Such lists would include information about individuals' characteristics, such as location, gender, and level of engagement, as well as contact information. Lists would also include both active participants and participants who completed their work with the activity, as well as participants who dropped out midway through implementation. Such ideal lists are rarely available, however, especially for indirect beneficiaries.
2. ETs should plan to work with USAID and IPs to develop and subsequently validate lists that are as comprehensive as possible. A key element of this process will be to understand what types of potential respondents might be systematically missing from the list and what biases, if any, their absence could be introducing in the findings. ETs should disclose any absences and biases in the design and final reports. If a somewhat complete list cannot be compiled, it is acceptable to start with recommendations from the IP and USAID, but plan for adjusting the selection as more information becomes available.

[16] Flick, U. (2018). *Designing qualitative research*. Second Edition. Qualitative Research Kit Series. SAGE Publications.

3. *Stratification:* Divide the population of intended respondents into smaller groups (i.e., strata) using one or more characteristics. If following a representative selection approach, all strata would remain in the sample for the next step. If pursuing a purposive selection approach, only the strata with the desired characteristics (e.g., active participants or participants who dropped out) would be included in the sample. Stratification can inadvertently result in silencing marginalized voices. ETs should modify the stratification strategy as needed to ensure the inclusion of such voices when appropriate.

4. *Respondent selection:* Within each relevant stratum, include all respondents if the stratum is small or randomly select respondents if the stratum is large.

2.5 To obtain breadth and depth, consider conducting a survey followed by random or purposive selection of respondents for more in-depth interviewing. This is a particularly attractive strategy if the target population of interest is large and resources are sufficient or if the ET has a list of program participants with contact information but no additional information to inform selection. Conducting a survey can shed light on the distribution of experiences and views in the population. Subsequent qualitative interviews with a random or purposive sample of survey respondents can deepen the ET's understanding of these experiences and views. It is best practice to inform survey respondents of future interview possibilities and ask them if they would like to opt in. If an evaluation will include a survey to inform subsequent interviews, ETs should plan for additional fieldwork time. Conducting a web-based or phone-based survey would require an additional four to six weeks of data collection, while a face-to-face, complex sample survey would require an additional twelve weeks minimum.

2.6 Upon completing respondent selection, validate the selection against the EQs and ensure that the information to be collected is sufficient to address them. This validation process will benefit from recognizing the types of information that respondents can be expected to contribute. Respondents can provide information about implementation processes, inputs, outputs, and outcomes. This information is usually well-suited to address EQs related to implementation processes and the mechanisms whereby an activity produces change (i.e., causal mechanisms). Respondents can also provide rich information about their own experiences and views about the changes they have experienced. ETs can usually aspire to capture the range of experiences and views, but not to draw conclusions about the distribution of experiences and views in the population or about program impacts.

When validating key informant and other respondent selections, ETs should ensure that more than one informant or respondent can speak to any given issue, allowing them to triangulate information across interviews. As discussed in Paint Pint 6, "Evidentiary support for statements," triangulation across respondents—as well as across data sources—is essential to increase confidence in evaluation findings.

2.7 Provide detailed information on respondent selection in the report, including any deviations from original respondent selection plans and their analytical implications. ETs should provide a general overview of the respondent selection approach in the main body of the evaluation report and include detailed information in an annex. Explanations should differentiate between key informants and other respondents, as well as provide selection information for each data collection approach (i.e., interviews, group interviews, and FGDs). They should also define what the population is and how the selected sample relates to the population, especially in the case of program participants and indirect beneficiaries. The report should also explain any deviations from the original respondent selection in response to logistical or security considerations, including the inability to contact intended respondents and refusals to participate, as well as the implications of these deviations on findings. ETs should report response rates, and ideally, contact and refusal rates, especially for program participants and indirect beneficiaries. See Box 2.3 for a definition of these rates. ETs should also reflect on and disclose any potential implications of nonresponse in the analysis. This should be noted in the limitations section as well as when presenting those findings that are impacted by the

missing information. As mentioned in Pain Point 1, "Case or site selection for small-n studies," ETs should use outbriefs as an opportunity to share information about the outcomes of fieldwork and any adjustments that were needed to the original respondent selection and outreach strategies.

Box 2.3: Reporting contact, refusal, and response rates

The contact rate is the proportion of potential respondents who answer to contact attempts (regardless of whether they ultimately participate in an interview) to all potential respondents. The refusal rate is the proportion of potential respondents who did not agree to participate to all potential respondents. Lastly, the response rate is the proportion of respondents who participated in an interview to all potential respondents.

When appropriate, ETs should rely on USAID and IPs for introductions to increase the likelihood of respondent cooperation while making clear that the evaluation is independent and participation is voluntary (and that provided information will be adequately protected). However, the communications inviting intended respondents to participate in interviews (and surveys) should come directly from the ET. When possible, ETs should also plan for multiple contact attempts and schedule calls to reach as many initially selected potential respondents as possible.

GUIDANCE FOR STUDY COMMISSIONERS

1. Prepare for the PE by obtaining from the IP the needed information for respondent selection to share with the ET together with other project documentation (e.g., annual and quarterly reports). The minimum required information should be a list of all program participants, including characteristics, such as location, gender, and level of engagement, as well as contact information.
2. Given that USAID often knows that it is going to conduct an evaluation of an activity at the design stage, it can take several steps to prepare for an evaluation from the outset. For example, if it is desirable for the evaluation to interview (or survey) program participants, consider requiring IPs to collect a participant list (and an indirect beneficiary list when appropriate). This can be done in ways that protect personally identifiable information and other sensitive information. Develop guidelines and training materials that IPs can use to ensure the quality of such lists.
3. Absent a pre-existing list of program participants, when preparing the evaluation SOW, include sufficient time and resources for ETs to work with USAID and IPs in developing lists of potential respondents, particularly of program participants and indirect beneficiaries.
4. Include a section in the SOW that identifies what information IPs will provide to ETs and how IPs, USAID, and ETs will collaborate to identify potential respondents at the evaluation planning stage.
5. Build flexibility into the evaluation timeline to ensure that ETs can adequately respond to fieldwork issues, like respondent unavailability and mobilization challenges. This will increase the likelihood that the respondent selection strategy is not compromised.

3. SOCIAL DESIRABILITY BIAS

SUMMARY

The Challenge: If an ET asks program participants if they are satisfied with a program or feel that it had a positive impact, many people will naturally answer "yes" regardless of their true perception. Despite this clear limitation, ETs frequently lean on these kinds of questions and risk drawing incorrect conclusions.

Minimal Guidance:

- Recognize the difference between mitigation and resolution of SDB (3.1).
- Identify and use sources of data and data collection methods that are less subject to SDB (3.2).
- In developing instruments, use indirect questioning, be thoughtful about question wording, and consider prefacing questions (3.4).
- During the interview itself, create a trusting atmosphere, look for cues of SDB, flag bias risks in notes, and follow up and probe (3.6).
- Take SDB seriously in analyzing data, report writing, and quality assurance and caveat findings accordingly (3.7).

THE CHALLENGE

SDB is the tendency to underreport socially undesirable attitudes and behaviors and to overreport more desirable attributes and behaviors. In the case of evaluation work, SDB most commonly manifests as interview respondents overstating the value, effectiveness, or satisfaction with the activity under evaluation. It is an unavoidable challenge in performance evaluations that depend on self-reports from program implementers and participants, but ETs typically lack the ability to measure it effectively.

There are diverse reasons why respondents might overstate positive views of an activity.

- They might want to simply be nice.
- Respondents might want to express their thanks to donors or implementers by presenting their efforts in a positive light.
- They might be answering strategically with the hope that it will lead to future benefits or funding.
- A related form of bias arises from "self-deception," whereby the respondent provides an overly positive assessment of a program but based on views of which they are convinced.[17]

This bias in reporting correlates with and mutually reinforces other forms of evaluation bias. For example, those who have more positive views of a program might be more likely to end up participating in an evaluation interview or FGD either through self-selection or donor/implementer selection. In a related vein, an evaluator faces certain incentives to tell donor clients what they want to hear (e.g., easier review process, positive feedback, future work), and positively biased findings make this more likely.

To some extent, the problem of SDB is recognized in PEs. Per ADS guidance, nearly all of the coded LER PEs (95 percent, 56 of 59) included a limitations section and almost half (47 percent, 27 of 57 that could be coded) listed some form of SDB as a limitation. Of those who mentioned SDB, 14 of 27

[17] See for a list of motivations: Bergen, N., & Labonté, R. (2020). "Everything is perfect, and we have no problems": Detecting and limiting SDB in qualitative research. *Qualitative Health Research, 30*(5), 783-792.

referred to it by name, while the remaining 13 of 27 referred to it by concept. Furthermore, of these PEs, most, 21 of 27 listed a mitigation strategy. Unfortunately, recognition of the risk of SDB in the limitations section does not generally appear to influence the interpretation of the findings. The concept reappeared in the findings, conclusions, or recommendations sections in only 5 of the 14 PEs that mentioned SDB by name in the limitations. Of these, most (four of five) only mentioned social desirability in one additional place. This suggests that ETs need to take SDB more seriously in the analysis and interpretation of findings and not limit it to a checkbox in a limitations section.

GUIDANCE FOR ETs AND LPs

3.1 Recognize the difference between mitigation and resolution of SDB. As noted above, many PEs identify SDB as a study limitation and note mitigation methods; however, they only rarely acknowledge this form of bias in the body of the report itself when interpreting the data. When a study is highly dependent on interviews, group interviews, and FGDs, it is important that ETs recognize that mitigation measures *mitigate* but do not *eliminate* the problem of SDB. As such, the importance of SDB should not be confined to a mention in a report's limitations section. If data is presented that could be subject to SDB, then this should be noted in the interpretation.

3.2 Identify and use sources of data and data collection methods that are less subject to SDB. It is important for ETs to not default to or be entirely reliant on KIIs, group interviews, and FGDs with individuals involved in the activity. Teams should look for potentially more objective data sources (e.g., monitoring data, administrative data, third-party surveys, or metrics) and look for respondents less subject to SDB. This might include third parties not involved with an activity but knowledgeable of it, including other donors, other CSOs operating in the space, journalists, and academics. Teams should lean on local team members to help identify these individuals.

When the team is reliant on interviews, teams should be thoughtful about whether individual interviews, small group interviews, or FGDs are more appropriate to mitigate SDB. For example, if there is a strong power imbalance between the potential interviewer and the interviewee, then interviewees might be more forthcoming surrounded by and in discussion with peers. Furthermore, more participatory data collection methods (e.g., community and process mapping), if well facilitated, offer a format potentially more conducive to building trust and eliciting honest responses than shorter interviews. Similarly, teams should consider whether interviews led by a local researcher or an international team member will produce more honest responses among a given population.

3.3 Limit access to the data, including donor access, and communicate clearly to interviewees about how the data will be used. While some donors are increasingly requesting access to notes and transcripts, the study team disagrees with this approach. Interview data should be treated confidentially, and the confidentiality of the interview should be stated upfront, along with a request for honest responses to questions. If interviewees understand that no one beyond the research team will know how they responded, this can reduce certain forms of SDB, in addition to reducing risk to study participants. While, in general, the study team is in favor of recording interviews (see Pain Point 4, "Qualitative data capture"), one reason not to record would be a high risk of strategic SDB (i.e., respondents are afraid to be on the record making negative comments).

3.4 In developing instruments, use indirect questioning, be thoughtful about question wording, and consider prefacing questions. Teams should review all instruments with an eye toward bias risks and how they can be minimized. First, qualitative and quantitative instruments should avoid direct questions likely to be subject to SDB. This includes direct questions about program effectiveness, program impact, or sensitive program challenges. For example, rather than asking training participants if they attended and participated actively in all sessions of a training program—something they might overstate in an interview or survey—interviewers can ask them about *other* participants' attendance and participation.

Second, teams should consider the most appropriate wording of questions. For example, rather than asking respondents about negatively framed "weaknesses" of an activity, an interviewer can employ a more positive framing and ask how the activity could be improved going forward.[18]

Third, teams should consider prefacing options. This might entail introducing a line of questioning with a prompt that recognizes challenges and complexity and is designed to encourage more honest responses. For example, an interviewer might state, "All programs confront challenges, and I want to ask you about some common challenges and see if you have experienced any of these." By noting that these are common challenges, it makes it easier for the respondent to acknowledge that these challenges exist in their program. In a similar vein, the interviewer can note at the appropriate point in the interview—while making eye contact—that they have a few sensitive questions that they want to ask, acknowledge the potential SDB, and remind the respondent that the interviews are confidential.

3.5 Pretest different approaches to questions subject to SDB. Pretesting and piloting are difficult to do in the context of a traditional PE where timelines are short, and data collection is concentrated in a short period of time. However, if prioritized, this can be built into the timeline and be led by local team members prior to the main data collection period. For example, one interviewed team leader reported regularly having locally based colleagues conduct pilot interviews. Using a cognitive testing approach, the interviewer can ask the question and then ask what the respondent understood by the question and how they arrived at their answers.

3.6 During the interview itself, create a trusting atmosphere, look for cues of SDB, flag bias risks in notes, and follow up and probe. Building trust and rapport in an interview setting is an important skill for qualitative researchers. There are a number of measures that can be taken, starting with crafting language for initial outreach, scheduling the interview, ensuring a suitable interview location, considering how the evaluation is presented in consent statements, using introductions to develop a sense of trust, and maintaining an open and engaged manner and posture during the interview itself. It is essential that interviewers set a tone that they are looking for frank and honest responses, emphasize their independence as evaluators, clarify how the information provided will be used, and highlight confidentiality measures.[19] This is particularly challenging in increasingly common remote interviewing.

In group interviews and FGDs, creating this atmosphere requires additional measures, including skilled facilitation and warm-up questions designed to encourage interaction. One team leader reported having participants go around the room and identify their favorite superhero, which was then used as a pseudonym throughout the interview. The exercise lightens up the session, builds rapport among participants, and—assuming they do not know each other—creates a sense of anonymity.

Interviewers and notetakers should be on the lookout for potential SDB. This might be indicated by the denial of any problems, challenges, or shortcomings; partial or vague answers; excessive and repeated praise for an activity, implementers, or donors; or nervous body language.[20] These should be flagged in the interview notes (e.g., writing "[SDB]" in brackets) to be taken into account during analysis.

When presented with potential SDB, interviewers can ask follow-up questions to probe for more detail, ask clarifying questions, request examples, or politely challenge responses. For example, an

[18] Direct question wording is sometimes incentivized by EQs and donor review of instruments. If ETs have to address so many EQs they have less flexibility in using interview time to build trust and ask indirect questions. Similarly, donor review might unintentionally encourage direct interview questions easily tied to the EQs. ETs might need to push back on questions that cannot be accurately answered and to add notes to submitted instruments to clarify the interview approach.

[19] While Pain Point 4, "Qualitative data capture," discusses the desirability of typing notes, it is important that the lead interviewers be able to maintain eye contact and create a trusting interview environment.

[20] Bergen, N., & Labonté, R. (2020). "Everything is perfect, and we have no problems": Detecting and limiting SDB in qualitative research. *Qualitative Health Research, 30*(5), 783-792.

interview might follow up with, "It sounds like you found the approach to be successful. Could you give me some specific examples?"

3.7 Take SDB seriously in analyzing data, report writing, and quality assurance and caveat findings accordingly. Teams are not required to blindly repeat in reports what was told to them in interviews. Instead, they should weigh the balance of the evidence in support of the findings. By triangulating among data sources, teams can identify if a given source (e.g., donor interviewees) is more positive (or negative) about results than other data sources. This includes comparing across interviewee types as well as between interviews and other data sources. Flagging potential SDB in notes, as suggested above, can aid in this process, and triangulation is particularly important where bias is suspected. Potentially biased sources of data can either be discounted or caveated. If done haphazardly, this discretion can undermine the systematic and replicable nature of qualitative research, so teams should take steps to document such decisions. Furthermore, "red team" or quality assurance checklists for analysis plans or draft reports should include and flag potential SDB.

GUIDANCE FOR STUDY COMMISSIONERS

1. Consider the risk of SDB in developing the SOW and EQs and in reviewing interview instruments and other deliverables. Be conscious of any personal potential confirmation biases and desires to see USAID and its program reflected in a positive light.

5. Do not request raw data. Evaluation data has the potential to be a public good and to be made publicly available. Doing so with qualitative data, however, requires disclosure to participants, requires resource-intensive anonymization efforts, and may impact the honesty of respondents. If qualitative data is to be shared, it should only be done with highly structured qualitative data collection approaches, bias concerns should be considered, and this should be planned and budgeted for from the beginning and disclosed to participants.

4. QUALITATIVE DATA CAPTURE AND MANAGEMENT

SUMMARY

The Challenge: Interview and FGD notes are the data that findings and conclusions should be derived from, and yet in some cases, the data provided to research teams might not be well captured or stored to allow for meaningful analysis.

Minimum Guidance:

- Use a systematic method to capture and organize data that is clearly articulated in the work plan (and note any deviations in the results report). Have a plan (including staffing) for consistently capturing a close to verbatim record of each qualitative event, either as a primary data capture strategy (if transcription is not possible) or a backup strategy (in the case of refusal to record) (4.1).
- Follow guidelines for informed consent and data protection (4.3).
- Practice regular recording of qualitative events if consent is given, as a backup to avoid data loss (absent extenuating circumstances that prohibit it) (4.5).
- LPs should provide regular, direct quality oversight over data capture (4.6).

THE CHALLENGE

The quality of data capture currently varies greatly across evaluations/teams, subcontractors/LPs, and locations. While the exact prevalence of practices is unclear, interviews and the PE report review revealed that some evaluators rely primarily on handwritten notes that may not be digitized, while other teams use standardized or structured digital note formats, and there is a wide variety of practices between these two poles. Poor quality (or absent) notes can lead to reports where findings are not substantiated (because the substantiating data is not captured), leading to commissioner doubt about the credibility of findings. It may be necessary to devote a lot of resources on the backend to redo the analysis, if that is even possible. There is also little consistency in the information provided in the evaluation reports about data capture. For example, only 32 percent of LER PE reports coded mentioned obtaining transcripts of qualitative events for analysis purposes, and overwhelmingly (95 percent, 56 of 59), there was no mention of systematic note-taking (e.g., templates, digital files, shared files). Due to the lack of information about plans for data capture in written deliverables (work plan/evaluation design and evaluation report), USAID has little visibility into this process.

Note-taking is the most common qualitative data capture strategy, but it has several challenges/limitations. A poor note-taking strategy can take away from the quality of the qualitative event (e.g., meticulous note-taking can distract from the conversation).[21] Notes may also lack important information for data interpretation. Some notes are not sufficiently comprehensive (e.g., probing or important contextual information might not be captured). Note-taking preferences also vary by team lead and team members. Some team leads would like the whole conversation captured verbatim (closer to a transcript), and others want paraphrasing.

Note-taking is also time-consuming to do well. Significant time is required for refinement and data management (e.g., cleaning up notes after the event and summarizing key takeaways and points for follow-up). During data collection, there may not be enough hours in the day to get everything done. These daily data processing tasks often slip during data collection due to these time limitations, leading to long delays in the circulation of notes to LPs or non-traveling team members.

[21] Morra Imas, L. G. & Rist, R. C. (2009). *The road to results: Designing and conducting effective development evaluations*. World Bank Publications.

The audio recording of qualitative events facilitates transcription and rigorous analysis, but recording is not always possible, and transcription can be time-consuming and costly. Additionally, there is often a trade-off between the desire to audio (or video) record a qualitative event to capture a record of the conversation and the risks associated with a recording (e.g., the potential of a data breach or that respondents will be less forthcoming).

Issues of consent and ethical research practices arise generally when considering data capture options due to the need to minimize the risks for participants and the research team members. There is a heightened risk for participants and team members, especially local team members, if research materials are seized in conflict areas or more restrictive environments for DRG work, and this may impact the ability to capture a record of qualitative events through notes or recording. Ethical issues also arise upon requests for the sharing of qualitative data, including with commissioners. This is because even with key identifiers (e.g., names, job titles) removed, it could still be possible to identify a participant based on a constellation of variables. Additionally, it would be impossible to de-identify KIIs targeting respondents based on their positions held.

GUIDANCE FOR ETs AND LPs

To recognize the variety of preferences and systems for data capture, overall, it is necessary to have minimum standards, but it is also important to give ETs some autonomy as to what works best for them and what is appropriate for the assignment and setting.

4.1 Develop and implement a data capture plan to consistently capture a verbatim or close-to-verbatim record of each qualitative event. The overall design for the study, including the later stages of analyzing and reporting, should be planned before the interviewing begins. The data capture plan should align with analysis and reporting needs and detail a strategy (including staffing) for note-taking, either as a primary data capture strategy (if transcription is not possible) or a backup strategy (if recording and transcribing). This plan should also include information on how team members will be trained to capture notes, how quality assurance is done on notes, whether interviews are recorded, whether transcriptions are produced, how confidentiality is protected, and the language of notes/transcripts. A good data capture plan should ensure sufficient redundancy in the process (e.g., backup recording and secondary notes) so that data is not lost if something is missed in the primary notes. Validate the data capture plan internally and by sharing it with the client/commissioner as part of the work plan. The results report methods annex should note any later deviations from the data capture plan.

4.2 Find the right size in the mix of skills on the ET. Ideally, all members of the ET involved in qualitative data collection should have training or experience in qualitative research. If this is not possible, at minimum, the team lead and evaluation specialist should have this training to ensure that qualitative data capture follows minimum standards and facilitates rigorous analysis. Avoid interpretation and the need to translate transcripts and, when possible, have interviews in the interviewees' language by staffing evaluations with speakers of the main languages that will be required.

It is also important for evaluations to find balance in team structure to benefit from local experience. Evaluation staffing also increasingly reflects localization in terms of team structure. LPs are increasingly deploying entirely local ETs, which can have language benefits but bring challenges in terms of ensuring adequate qualifications in qualitative research. Oftentimes, a remote team member will support information gathering, analysis, and writing of the report to USAID standards.

Teams should have at least one secondary interviewer in each interview who serves as the primary notetaker. Ideally, this should be an ET member serving in this role (team lead, local expert, or mid-level evaluation specialist), and team members should rotate roles across interviews (alternating lead interviewer and notetaker). If using a dedicated notetaker outside the core ET, it is important to validate necessary skills (e.g., through experience or even a typing test) and provide adequate training and oversight to ensure that notetakers understand the purpose of the research and are able to

faithfully capture information from the data collection events. If interviews are not audio recorded, teams should use two notetakers (i.e., the primary and secondary interviewers both take notes) to have a backup and allow comparison/reconciliation between the versions.

4.3 Follow guidelines for informed consent and data protection, minimizing the collection of and access to raw or identifiable data. The guiding principle for ETs should be to make and follow a data capture plan that protects the interviewee and interviewer as much as possible. Consent statements should cover recording and direct quoting in the report.[22] It is important to remember that most institutional review boards require consent protocols to outline who will have access to the data. If a commissioner requests the data after the data collection has taken place, this would not have been in the consent statement and is, therefore, not something that the participants agreed to. For this reason and because the sharing of raw qualitative data will likely have a chilling effect on what is shared with ETs, as noted in Pain Point 3, "Social desirability bias," the sharing of raw or close-to-raw data with commissioners is not recommended. Teams can consider giving access to other processed (and de-identified) analysis products instead, such as codebooks, etc.

LPs (and the teams they manage) should also follow security protocols[23] to protect the data collected during PEs: limiting the collection of and access to personally identifiable information (e.g., by labeling data files by code instead of respondent names), practicing secure storage and timely transfer of notes and recordings (including using password-protected devices and prompt uploading to cloud services and deleting from local devices), and destroying notes and recordings after the conclusion of the evaluation.

4.4 Encourage teams to follow good practices for note-taking. Recognize that the approach for note-taking may need to adjust based on whether the assignment is an evaluation or assessment and whether the interview is formal or semi-formal/casual (e.g., for background information, etc.). Create a note-taking template prior to data collection and use this systematically. This should contain consent language, interview questions, metadata, and a comments section on the interview. Common metadata includes the date, interviewers, interviewees, type of interview, respondent type, location, field notes relevant to the interview, linked documents, initial ideas for analysis, and anonymizing references, generally a respondent code.[24] The template should also include a system to document additional relevant information about the qualitative event, such as anything noteworthy about the interview (e.g., the interview was cut short after 20 minutes, lack of privacy, etc.) and interviewer observations (e.g., body language, potential SDB in responses, etc.).[25] The team developed a sample note-taking template in Appendix 3: Note-taking Guidance.

During the interview, the lead interviewer should focus on engagement with the interviewee. The secondary interviewer/primary notetaker should take detailed, close-to-verbatim notes on a small laptop. If it is not possible to take a backup audio recording because of the cultural or security context, then both team interviewers should endeavor to take good notes and minimize distractions from the conversation by maintaining eye contact while writing.[26]

After the interview, but within 24 hours, the secondary interviewer should clean the notes, and the lead interviewer should review them. This review and timeframe are particularly important if the

[22] In sensitive situations, consider incorporating multiple opportunities for informed consent (e.g., asking again at the end of the interview to ensure the interviewee is still comfortable with using their responses for analysis or if there is anything they would prefer not to be incorporated) and sharing notes with the respondent to check the recollection of responses and agree on what was said (Morra Imas, L. G. & Rist, R. C. (2009). *The road to results: Designing and conducting effective development evaluations*. World Bank Publications.). For the full USAID general requirements on informed consent, refer to CFR 225.116.

[23] For a broader discussion of data protection considerations and practices, refer to USAID's Using Data Responsibly guidance.

[24] Gibbs, G. R. (2021). *Analyzing qualitative data*. Sage Research Methods. SAGE Publications.

[25] Morra Imas, L. G. & Rist, R. C. (2009). *The road to results: Designing and conducting effective development evaluations*. World Bank Publications.

[26] Morra Imas, L. G. & Rist, R. C. (2009). *The road to results: Designing and conducting effective development evaluations*. World Bank Publications.

secondary interviewer lacks subject matter or contextual expertise. Besides reading through the notes to fill gaps and make corrections, it is important to add immediate interviewer interpretations, thoughts, and reactions. It is also important for interviewers to document the level of interviewer confidence in responses using a standardized system.[27] Time needs to be accounted for during data collection for this daily synthesis of notes (it may not be feasible to maximize the number of interviews daily and still complete these necessary processing tasks).

Handwritten notes should not be the primary data capture strategy unless required for security reasons or extreme respondent sensitivity (and even then, notes should be digitized later, if possible). If note-taking on a laptop is not possible, notes should be taken by hand by both interviewers and then digitized, reviewed, and securely stored within 24 hours.

4.5 Absent extenuating political or security circumstances, practice regular recording of qualitative events if consent is given and recording is unlikely to undermine frank responses. Such recordings serve as a valuable backup to avoid data loss from note-taking errors. Though the strategy used in each evaluation will vary depending on the nature of the data collection and specific context, in many contexts, teams should be able to easily record qualitative events using a password-protected phone or small audio recorder to capture a backup record of what was said that could be used to supplement notes.[28] Teams should also capture notes during and after each interview, even if the interviews are recorded, to ensure that data is not lost if a recording is lost or portions of it are unintelligible.[29]

The time and budget for a "traditional" PE may not allow for transcription (and, if needed, translation) of recordings, but if it is possible, some teams may choose to transcribe all recordings for analysis. When planning transcription resources, teams should allocate three to five hours for each one hour of interviews. Artificial intelligence (AI)-powered transcription technologies may greatly reduce the time needed for this work, if privacy and security concerns are addressed. (Box 4.1).

Box 4.1: AI-Powered Audio Transcription Options

1. **For virtual interviews and discussions:** Native transcription features are now common in virtual meeting platforms, such as Google Meet, Zoom, and Microsoft Teams. However, a key consideration in using AI for transcription services for USAID is the requirement to obtain approval for recording meetings.
2. **For in-person interviews and discussions:** If using a laptop, Microsoft Word's dictate function may sufficiently capture the conversation. Teams can also upload externally captured audio recordings to be transcribed by an automatic web-based service. These include Express Scribe (free), GoodNote, Rev, F4transkript, and F5transkript. Before selecting a service, teams should confirm that the data is stored securely and will not be used for algorithm training.

4.6 Select LP staff should have access to all of their team's qualitative data and provide regular quality oversight. Notes should be made available to select LP staff responsible for oversight on a daily basis as they are finalized, and LPs should spot-check notes for timeliness and

[27] For example, note-takers could write "CHECK" in the notes or use a color-coding scheme to indicate a point that should be verified or might use brackets "[ ]" to offset interjections and notes.

[28] Even if consent is given, interviewers should look for signs that the recording might be affecting responses (e.g., by observing if interviewees regularly look at the recorder or asking a modified version of an earlier question after the recorder has been turned off at the end of an interview).

[29] Morra Imas, L. G. & Rist, R. C. (2009). *The road to results: Designing and conducting effective development evaluations*. World Bank Publications.

completeness. At least two personnel from the LP should have clearance to access identified data for this purpose. One interviewee noted that consultant contracts can be a useful tool to ensure compliance with data standards. For example, payment for data collection can be tied to the delivery of complete notes.

GUIDANCE FOR STUDY COMMISSIONERS

1. Commissioners should communicate expectations about data capture practices (e.g., transcription) during evaluation SOW development so that evaluations can be appropriately budgeted and staffed.

2. Commissioners should be sensitive to the ethical issues that arise during data collection: the general need for all users and uses of the research to be outlined in the informed consent (limiting subsequent deviations) and the need for the protection of study participants. This requires limiting access to raw or identifiable data to members of the research team. If transcriptions are to be provided as a deliverable, then this should only be done in specific circumstances where transcripts can be meaningfully anonymized and when commissioner access to transcripts outweighs the risk of participants being less forthcoming during interviews. Furthermore, time and funds need to be budgeted for anonymization.

5. QUALITATIVE DATA ANALYSIS

SUMMARY

The Challenge: Without a systematic, documented, and somewhat replicable approach to data analysis, research teams risk several forms of bias, including confirmation bias and deriving conclusions from early interviews, recent interviews, or dynamic and memorable interviewees. In addition, it becomes difficult for teams to collaborate, conduct quality control, or revisit findings and conclusions that are poorly documented.

Minimum Guidance:

- Employ a systematic process for arriving at findings and documenting that process. The work plan should outline: 1) the types of inquiry and analysis strategy that will be used; 2) whether notes or transcripts will be analyzed, how they will be analyzed, and by whom; and 3) the procedure for documenting how teams will confirm findings collectively and establish inter-coder reliability (5.1).
- LPs and team leads should provide leadership and oversight throughout the evaluation to ensure that the analysis plan is carried out faithfully (5.5).
- A minimum of two team members (with one core team member) should be involved in coding (5.6).
- Combine systematic analysis of qualitative data with other sources (e.g., desk review, surveys, analysis of secondary data) to triangulate findings (5.8).

THE CHALLENGE

Due to the short-term nature of traditional PE assignments, PE teams do not always employ systematic analysis methods to arrive at findings. Not using a systematic method, such as coding, leaves room for subjectivity and opens the door for real or perceived cherry-picking of data to support preconceived hypotheses.

ETs may lack the time, labor, capacity, or buy-in to comply with a proscribed analysis process. Specifically, team members included for their subject matter expertise may not have had training or experience with qualitative methods, and team members need to buy into and fully participate in the analysis process (e.g., accepting findings that contradict their priors) for it to work as intended.

There are also a number of pitfalls/challenges that can diminish the usefulness of a systematic analysis approach. If completed by untrained team members, information can be lost (e.g., not coded) or incorrect. Coding schemes that are too ambitious can also be problematic. If they are not practical, considering the time allotted to analysis, the team may run out of time and rush or fail to complete parts. Additionally, the focus of semi-structured interviews is often tailored to the specific interviewee and follows the thread of the conversation. If evaluations rely heavily on these types of interviews, elaborate coding to allow the comparison of responses across interviews may not be a good use of time. Three USAID and team lead interviewees noted a common failure to adequately incorporate document reviews or monitoring data into the final analysis.

Even when systematic analysis methods are used, they are often not documented to an extent that allows the reader to understand how findings were derived or to facilitate replicability. In the PE report review, there was a nearly even divide between reports that explained how the qualitative data was analyzed and those that did not ("no" at 49 percent, 29 of 59; "yes" at 51 percent, 30 of 59). LPs may not have access to teams' analysis files, inhibiting quality oversight.

According to the PE report review, of those reports that provided information about analysis methods, the majority (57 percent) fell into the category of light-touch thematic coding. This approach involved strategies like identifying key themes and patterns and employing techniques like outcome mapping and rapid/thematic analysis with Excel. Twenty-seven percent (8 of 30) of the reports showed evidence

of rigorous coding in their qualitative analyses (e.g., preset and emergent codes using specialized software). Of these answers, MAXQDA was the most used software, followed by Dedoose. Interviews echoed that the dominant strategy to organize qualitative data for systematic analysis is coding notes or transcripts, but some teams do not use systematic strategies to organize and analyze qualitative data.

GUIDANCE FOR ETs AND LPs

The goal of qualitative analysis in evaluations is to: 1) make sure that what the ET reports about its data is defensible and clearly linked to the data and 2) capture the universe of themes, ideas, opinions, etc., present in the data, as recurrent and divergent themes, ideas, and opinions increase the ET's understanding of the activity under evaluation. According to USAID's Evaluation Policy, the analysis method applied should:1) "ensure, to the maximum extent possible, that if a different, well-qualified evaluator were to undertake the same evaluation, he or she would arrive at the same or similar findings and conclusions;" and 2) "use to the maximum extent possible […] social science methods and tools that reduce the need for evaluator-specific judgments."30 In the discipline of political science, these aims have been framed around a desire for transparency in qualitative research that is comparable to standard practices for transparency and reproducibility in quantitative research).31

5.1 Employ a documented systematic approach for arriving at findings from all data sources, including, at a minimum, structured thematic or content analysis for qualitative data. Purpose should guide the analysis strategy. The analysis should focus on the primary EQs and be reasonable enough to be completed in the time allotted.

The evaluation work plan/design should outline the types of inquiry and analysis strategies (Box 5.1) that will be used to arrive at findings. This description should clearly state whether notes or transcripts will be analyzed, how they will be analyzed, and by whom. The evaluator should make clear in the methodology section of the work plan whether and how the data will be coded as part of the analysis process and whether and how this coded data will be included in the final report.[32] The work plan/design should include a procedure for documenting how teams will confirm findings collectively and establish inter-coder reliability.[33]

Box 5.1: Qualitative Analysis Strategies

The three core qualitative analytical strategies are inductive (explanation building), deductive (pattern matching), and abductive (continuously developing and testing theories).[34] Most current traditional PE analysis is inductive/explanation building. To weave in deductive and abductive inquiry throughout the process, as feasible, allows the data collection to be more focused and increases confidence in findings by demonstrating that alternative explanations were considered and rejected.

5.2 Systematic analysis of qualitative data in PEs will most often be facilitated by first coding what was discussed into coherent categories (preset and emergent).[35] Coding involves

[30] USAID. (2020). *USAID evaluation policy*. https://www.usaid.gov/sites/default/files/2022-05/Evaluation_Policy_Update_OCT2020_Final.pdf

[31] Kapiszewski, D., & Karcher, S. (2021). Transparency in practice in qualitative research. *PS: Political Science & Politics*, *54*(2), 285-291. https://doi.org/10.1017/S1049096520000955

[32] USAID. (2013c). *Focus group interviews*. [Technical note]. Version 1.0. Monitoring and Evaluation Series.

[33] USAID. (2023). *Assessment of study quality (ASQ) tool*.

[34] Brinkmann, S., & Kvale, S. (2019). *Doing interviews*. Sage Research Methods. SAGE Publications; US General Accounting Office. (1990). *Case study evaluations*. https://www.betterevaluation.org/sites/default/files/10_1_9.pdf; Patton, M. Q. (2002). *Qualitative evaluation checklist*. Evaluation checklist project. The Evaluation Center, Western Michigan University.

[35] Taylor-Powell, E., & Renner, M. (2003). *Analyzing qualitative data.* (G3658-12). Program Development and Evaluation. University of Wisconsin-Extension.

organizing qualitative data by labeling or categorizing passages or parts of transcripts or other data sources into a matrix or database so that the information can be readily retrieved, searched, compared, and contrasted.[36] Codes make it easier to compare data, identify any themes or patterns present in the data, and group data points by characteristics (e.g., type, geography, gender, etc.).

Once data is coded, researchers can identify patterns and connections within and between categories. These include:

- **Within category descriptions and between comparisons:** Similarities or differences in people's responses according to a demographic (e.g., gender, age, etc.) or other category (e.g., USAID/IP responses vs. program participants, location 1 vs. location 2 respondents, government respondents vs. citizen respondents, document review data vs. interview data).
- **Larger categories:** Work up from specific codes to larger ideas and concepts.
- **Relative importance:** Counts (when appropriate).
- **Relationships:** Themes that occur together, which might suggest cause and effect.

Coding is essential when there are more than a few (five or more) qualitative interviews to analyze together because it allows the ET to complete the analysis by interpreting and attaching meaning and significance to the full set of information collected.[37] While coding can facilitate analysis based on the frequency counts of words or themes, and it might be desirable to track frequencies in certain instances (e.g., answers to a structured tool with a large number of respondents), coding is an organizational tool that allows the team to review and synthesize data points holistically according to *any strategy*. In most cases, frequency counts should not be the primary goal of coding for traditional qualitative PEs.

5.3 Basic rapid analysis using manual thematic coding is sufficient in certain circumstances.[38] Based on the experience of the study team and many interviewees, organizing data according to codes or key themes in Microsoft Excel is a common practice. Generally, basic information about the qualitative event (e.g., type, location, date, gender, etc.) is input into a spreadsheet with key themes or observations, including illustrative quotes (see the template in Appendix 4: Rapid-Analysis Coding Matrix Template). In the final matrix, the data is organized with similar themes together to facilitate writing, and the team can perform basic filtering and sorting to compare responses by category. The advantages of this approach are that it is accessible (i.e., does not require specialized software or training) and does not require special data preparation before analysis. However, the types of tabulations and comparisons possible are more limited than what is available in dedicated coding software. This approach is most appropriate when there is a relatively small number of interviewees (less than 30), when very different questions are being asked to different interviewees using highly tailored questionnaires (in which case there are fewer benefits to software assistance, and you might need multiple small analyses), or when data quality capture is not good (e.g., no transcripts and variable note quality).

5.4 Some teams may want to pursue (and some Missions may request) coding using qualitative data analysis software. This can be helpful in finding common themes with a large

[36] USAID. (2013c). *Focus group interviews*. [Technical note]. Version 1.0. Monitoring and Evaluation Series.

[37] Taylor-Powell, E., & Renner, M. (2003). *Analyzing qualitative data.* (G3658-12). Program Development and Evaluation. University of Wisconsin-Extension.

[38] Gale, R. C., Wu, J., Erhardt, T., Bounthavong, M., Reardon, C. M., Damschroder, L. J., & Midboe, A. M. (2019). Comparison of rapid vs in-depth qualitative analytic methods from a process evaluation of academic detailing in the Veterans Health Administration. *Implementation Science, 14*(1), 11. https://doi.org/10.1186/s13012-019-0853-y; Neal, J. W., Neal, Z. P., VanDyke, E., & Kornbluh, M. (2015). Expediting the analysis of qualitative data in evaluation: A procedure for the rapid identification of themes from audio recordings (RITA). *American Journal of Evaluation, 36*(1), 118-132. https://doi.org/10.1177/1098214014536601; Vindrola-Padros, C., & Johnson, G. A. (2020). Rapid techniques in qualitative research: A critical review of the literature. *Qualitative Health Research, 30*(10), 1596-1604. https://doi.org/10.1177/1049732320921835

number (10 to 15 or more) of fully transcribed interviews.[39] Using software allows for retrieving all text coded with the same label and for analyzing further analytic questions, especially if lists of codes are developed as a hierarchy.[40] Although software facilitates or supports qualitative analysis, the coding and interpretation still rest with the evaluator.[41] Shared software can also give the LP management team continuous updates that help reassure from the management perspective that the analysis is proceeding according to plan. Software options include Dedoose, MAXQDA, NVivo, and ATLAS.ti.

However, there are logistical and staffing limitations to the use of qualitative analysis software that must be overcome. Preparing and entering data for analysis may be time-consuming, depending on the type of data and the software chosen, and require significant research assistant support.42 The coding work itself is also time-consuming. According to USAID guidance, "Even with the use of a software package, coding qualitative…data requires a substantial amount of time, and the evaluation manager should allow for adequate time in the SOW".43 Often, when this type of coding must take place on a compressed timeline, coding is delegated from "senior/core" ET members to support analysts, potentially leading to less precise results. Even when the budget and timeline permit core ET members to complete the coding, it can be hard to find a full team who knows how to use the same qualitative analysis software, particularly when the team is primarily local researchers. The budget for qualitative analysis software is also difficult if not using a free version of the software.44 Consultant team members likely will not have their own software licenses. LPs likely will have some licenses, but they may be procured through (and thus tied to) a project or difficult to share with consultant team members. Commissioners may not understand that software costs need to be included as other direct costs in evaluations.

5.5 LPs and team leads should provide leadership and oversight throughout the evaluation to ensure that the analysis plan is carried out faithfully. LPs should follow quality assurance protocols to ensure the analysis is rigorous in both timeliness and quality, oversee methodology and codebook development, ensure coders are adequately trained, and test that coders are coding transcripts in the same way, i.e., there is high inter-coder reliability.[45] The LP should be prepared to discuss and share example analysis matrices or documentation with the donor/USAID. Team leads should also provide leadership and oversight by ensuring that team members understand and follow analysis processes.

It is important to involve the full ET in analysis by creating space for people to critically reflect on findings and arrive at findings as close to reality as possible.46 Group preliminary analysis (often in preparation for the outbrief) is important because noting information while it is fresh helps avoid

[39] Balbach, E. (1999). *Using case studies to do program evaluation*. California Department of Health Services.; Gilbert, L. S., Jackson, K., & Di Gregorio, S. (2014). Tools for analyzing qualitative data: The history and relevance of qualitative data analysis software. In J. M. Spector, M. D. Merrill, J. Elen, & M. J. Bishop (Eds.), *Handbook of Research on Educational Communications and Technology* (pp. 221-236). Springer. https://doi.org/10.1007/978-1-4614-3185-5_18

[40] Gibbs, G. R. (2021). *Analyzing qualitative data*. Sage Research Methods. SAGE Publications.

[41] Vaessen, J., Lemire, S., & Befani, B. (2020). *Evaluation of international development interventions: An overview of approaches and methods*. Independent Evaluation Group, World Bank.

[42] Vaessen, J., Lemire, S., & Befani, B. (2020). *Evaluation of international development interventions: An overview of approaches and methods*. Independent Evaluation Group, World Bank.

[43] USAID. (2013c). Focus group interviews. [Technical note]. Version 1.0. Monitoring and Evaluation Series.

[44] Free versions of software are rare. For example, MAXQDA has a free trial, but requires purchasing a license for intensive use. While Dedoose once was free, it now requires a paid monthly subscription. Taguette and QualCoder have limited capabilities. Most other software requires the purchase of a license.

[45] According to O'Connor and Joffe (2020), 10–25 percent of transcripts should be double coded. There are a number of tests for inter-coder reliability, but for the purposes of a traditional PE, monitoring the percent agreement is likely sufficient. (O'Connor, C., & Joffe, H. (2020). Intercoder reliability in qualitative research: Debates and practical guidelines. *International Journal of Qualitative Methods, 19*.) Miles and Huberman (1994) suggest a standard of 80 percent agreement on 95 percent of codes as an acceptable threshold. If low reliability is found, teams can choose to increase the use of double coding or discard low-performing codes to address the issue. If kept, codes with low reliability need to be reconciled, either through consensus or the introduction of a third coder, and use either a "majority rules" or "final judgment" approach. (Miles, M. B., & Huberman, A. M. (1994). *Qualitative data analysis: An expanded sourcebook*. Sage Publications.)

[46] Kusters, C., van Vugt, S., Wigboldus, S., Williams, B., & Woodhill, J. (2017). *Making evaluations matter: A practical guide for evaluators*. Centre for Development Innovation, Wageningen University & Research.

recency bias. Teams should also hold periodic meetings (including the LP) to discuss coding and reflect on the coding frame and any necessary adjustments. Any significant disagreement in the findings/interpretation among team members should be discussed and, if not resolved, documented in the report.

5.6 Evaluations should be adequately budgeted and staffed to support systematic analysis. Qualitative analysis skills should be a requirement for the core ET roles (i.e., team lead, evaluation specialist). Research assistants are critical for supporting quality assurance processes as well as the analysis itself in some cases (see below). Research support and management will need to be adequately budgeted to ensure sufficient resources for robust quality assurance. Local team members should have core roles in data analysis and interpretation using an agreed systematic process with embedded local capacity-strengthening support. If an evaluation budget is not large enough to support adequate staffing to analyze a large volume of data, then commissioners should account for this in the approach by planning a more targeted evaluation that collects and carefully analyzes less data.

A minimum of two team members (with one core team member) should be involved in coding. Ideally, coding would be done by members of the team that conducted the fieldwork; however, this is often challenging given the timeline and level of effort considerations. While evaluation firms often employ junior researchers to code qualitative data, this can lead to major quality control issues if coders have inadequate experience or do not understand the context or the intervention. If coding is done by researchers not involved in fieldwork, then: 1) that coder should have strong experience coding; 2) they should be involved in the evaluation from the beginning of the study (e.g., supporting the desk review); 3) there should be in place a rigorous quality control process, whereby coders are trained on the codebook, undertake a pilot coding activity, use double coding for a set of notes/transcripts, and have their code regularly spot checked; and 4) this should be done when there are transcripts or close-to-transcription-quality notes rather than rough interview notes.

5.7 AI is an emerging tool that, with caution, could be used to speed up the coding process or help uncover hidden themes. The use of AI in qualitative analysis has the potential to offer powerful assistance to researchers. AI tools can code segments of text given predefined coding rules, which can then be validated against samples of human coding,[47] and allow researchers to focus on in-depth interpretation and development of rich narratives. Many AI tools (Box 5.2) facilitate efficient analysis by returning answers to detailed questions on transcript content with supporting quotes. They can also automatically generate summaries and thematic analyses of documents, which could be an efficient way to share qualitative analysis products with commissioners when requested. AI-assisted qualitative analysis may also be able to suggest initial codes for categorizing data and identifying emerging themes, though teams may be more comfortable keeping this framework development with core team members.

Box 5.2: AI-powered qualitative analysis tools

1. Qualitative analysis tools: AILYZE, Fathom AI (formerly Avalanche), CoLoop, AI-driven sentiment analysis (now within traditional qualitative analysis software such as ATLAS.ti), MAXQDA AI Assist, NVivo 15, HuggingChat, and flan R packages (which uses large language models)
2. Document and literature review tools: ChatGPT, QDA Miner, Rayyan, Google Gemini, and Elicit
3. Polling tools: Polis

[47] Gilardi, F., Alizadeh, M., & Kubli, M. (2023). ChatGPT outperforms crowd-workers for text-annotation tasks. *Proceedings of the National Academy of Sciences, 120*(30). https://doi.org/10.48550/ARXIV.2303.15056; Rathje, S., Mirea, D.-M., Sucholutsky, I., Marjieh, R., Robertson, C., & Van Bavel, J. J. (2023). GPT is an effective tool for multilingual psychological text analysis. *Proceedings of the National Academy of Sciences, 121*(34). https://doi.org/10.1073/pnas.2308950121

However, many of these algorithms are a "black box" (i.e., users do not know exactly how they arrive at their output, and they may "hallucinate" incorrect output), so they should not be relied on as the primary data analyst to arrive at evaluation findings. Rather, the literature suggests that AI should be used as part of a broader analytical strategy (e.g., "human–AI collaboration"). At a minimum, ETs should check any AI-generated output for accuracy and validate any AI-coded qualitative events against a subset of human-coded events. While some AI platforms store user data and use it to further train their models, others do not store data and adhere to industry-leading standards in data encryption while processing uploaded data. Researchers need to be careful to adhere to proper data security practices when using AI for qualitative analysis, including not uploading identifying respondent information to platforms that are not secure and fully understanding the data security protocols of AI systems. Planned use of these tools should be discussed in the evaluation work plan/design.

5.8 Combine systematic analysis of qualitative data with other sources (e.g., desk review, surveys, activity monitoring, evaluation and learning plan (AMELP) data, other secondary data) to triangulate findings. Together, with the systematic ruling out of alternative explanations and the explanation of "outlier" results, triangulation bolsters the validity of the findings by demonstrating the consistency of evidence across data sources. Particular strategies for making such comparisons, such as pattern matching, explanation building, and thematic review, can (but do not have to) be used.[48] Because not all sources will agree with one another, the ET should be transparent about the disagreements by sources and attempt to make sense out of inconsistencies.[49] Where possible, it is also important to demonstrate the consistency of findings across respondent types within data sources (e.g., different types of interviewees).

ETs should also apply a systematic strategy to document review (e.g., thematic coding). Excerpts from documents should be captured in a way that records and preserves context.[50]

GUIDANCE FOR STUDY COMMISSIONERS:

1. During planning, ensure that evaluations are adequately budgeted to allow a staffing structure that facilitates high-quality structured analysis. Beyond the core roles of the team lead, a local expert, and an evaluation specialist (if not covered within the two previous roles), LP headquarters support should include a manager/project coordinator and research assistant support.
2. If transcription and software coding are expected, this needs to be communicated at the SOW stage to ensure that recruitment prioritizes these skills. Depending on the final composition of the ET, the evaluation budget may need to be higher and the analysis time frame at least slightly longer. Generally, if the use of qualitative data analysis software is expected, the time frame to deliver the findings report will need to be extended by about two weeks (from four to six weeks) to allow for this.
3. Because the volume of data collected can become an issue for a time- or budget-constrained evaluation, in some circumstances, a better strategy is to collect less data and spend more time/resources on analysis.

[48] Bans-Akutey, A., & Tiimub, B. M. (2021). *Triangulation in research*. Academia Letters. https://doi.org/10.20935/AL3392; Lewis-Beck, M., Bryman, A., & Futing Liao, T. (2004). Triangulation. In M. Lewis-Beck, A. Bryman, & T. Futing Liao, *The SAGE Encyclopedia of Social Science Research Methods*. Sage Publications, Inc. https://doi.org/10.4135/9781412950589.n1031; Morra, L. G., & A. C. Friedlander. (1999). *Case study evaluations*. Number 2. Operations Evaluation Department (OED) working paper series. World Bank Group. http://documents.worldbank.org/curated/en/323981468753297361/Case-study-evaluations

[49] To avoid overly long reports, minority-dissenting views can be noted in footnotes.

[50] Patton, M. Q. (2002). *Qualitative evaluation checklist*. Evaluation checklist project. The Evaluation Center, Western Michigan University.

6. EVIDENTIARY SUPPORT FOR STATEMENTS

SUMMARY

The Challenge: There is conflicting guidance on the evidentiary support required for finding and conclusion statements. In some cases, tangential and poorly supported findings and conclusions are offered in reports, but proposed solutions to ensure evidentiary support often create other problems, such as requiring highly structured interviews, treating qualitative data like quantitative data, and ignoring many of the strengths of qualitative data.

Minimum Guidance:

- The source of evidence should be cited in a way that provides basic information about the source while still maintaining confidentiality (6.2).
- Findings must be based on multiple data points (6.3).
- Reports should be organized by findings and not divided by data source (6.4).

THE CHALLENGE

Interviewees expressed a few common concerns with qualitative evidence in typical PEs:

- Qualitative evidence, at times, comes off as anecdotal and unconvincing.[51]
- The writing frequently does not convey important information that readers want to know, including a sense of the frequency or commonality of a sentiment, whether findings are representative of a larger population, and the reliability of interview sources.[52]
- Findings are not linked to data sources or cited and are not triangulated based on multiple data sources, including desk and document reviews.[53]

While there is often agreement on these limitations, there is no consensus on what the appropriate evidentiary support for a given qualitative finding should be. Not surprisingly then, study users also have different expectations of what support is required, and donors and IPs are likely to question the empirical support for findings that they disagree with or might reflect poorly on them.

Part of the challenge is that many readers and users want qualitative data to look more like quantitative data to be confident in the findings. For example, these readers may want to see statements like "15 of 67 interviewees felt that a particular programmatic approach was ineffective." While some qualitative research can benefit from the logic and tools behind quantitative methods, this perspective risks ignoring the benefits of qualitative evidence and producing research that is neither good quantitative nor good qualitative research. Several interviewees noted that both researchers and users often do not adequately understand qualitative methods or the logic of qualitative inquiry.[54]

There are at least three strong reasons to be cautious in quantifying qualitative data:

1. In most cases, qualitative data comes from a sample that is not representative of a larger population, so percentages and even frequencies can be misleading.

[51] Interviews with two USAD staff (1, 2). This section of the report includes interview sources to illustrate point 6.2.

[52] Interviews and group interviews with USAID staff (1, 15), team lead (5), and firm/team lead (3). According to the PEs report review, 15 percent (9 of 59 PEs) used exclusively vague language, such as "interviews suggest;" 61 percent (36 of 59) used terms like "some interviewees", and an overlapping 66 percent (39 of 59) used more specific terms like "a plurality," "most," "one", or "two."

[53] Group interviews with USAID staff (1, 9, 15).

[54] Interviews and group interviews with team leaders (8, 14) and firms (9).

2. In a related vein, quantifying qualitative responses equally weights study participants, and some respondents might be more or less compelling sources of information. This might be because of their position, level of expertise, or willingness to speak honestly.

3. One of the advantages of qualitative research is its less structured nature and the ability to target different questions to different respondents. As such, in many cases, not every study participant is asked the same set of questions or asked about topics with comparable language.

GUIDANCE FOR ETs AND LPs

6.1 Quantify qualitative data only when using highly structured instruments on a large sample. Quantifying qualitative data makes sense if the study: 1) includes a large sample of respondents that aim to be representative of a larger population and 2) employs a highly structured interview instrument, where the same questions are asked of respondents in the same way. If the study does not meet these two criteria, researchers should not adopt a quantification approach. If a study commissioner has a strong preference for data presented in a quantified way, then such an approach needs to be determined at the design stage. Of course, there might be other options to balance the strengths of quantifiable and qualitative data. For example, even if a survey is not able to be conducted, a few high-priority, close-ended, mini-survey questions could be added to the end of a semi-structured interview. This might not address the representativeness challenge, but it would ensure that all respondents are asked the same questions in the same way.

6.2 The source of evidence should be cited in a way that provides basic information about the source while still maintaining confidentiality. This can be done:

- In the text: "In a group interview, one youth participant argued that…"
- Parenthetically: "(GI 6, youth)," where GI stands for group interview.
- As a footnote or endnote: "Group interviews with youth (6)."

Citing interviews in the same way as documents allows readers to know the source and extent of the evidence. Teams should also ensure that categories of interviewees reported in citations are adequately large so as to not undermine confidentiality. While this particular guidance document is not a research study, in this section, the study team has adopted the convention of citing interviews using footnotes to illustrate this practice. If interviews are not cited parenthetically or as footnotes, then narrations should include a greater level of specificity. For example, narrations should avoid phrases like "interviews suggest" in favor of more specific language like "a plurality of," "most," "one," "two," or "a minority of interviewees/participants suggest."

6.3 Findings must be based on multiple data points. The view of or quote from an individual respondent can offer valuable perspective, but it must be combined with other data points to make a summative finding. Noting, for example, that one program participant reported that a program component was valuable offers a potentially important data point, but this alone is insufficient evidence to serve as a finding about the utility of the component. This data point needs to be combined with other complementary data points to make a summative finding or conclusion, as shown in Box 6.1 below. Teams should take outlier viewpoints as factors to be further explored and validated in future interviews, through other sources, or through additional research. Of course, individual viewpoints can be very valuable in identifying good ideas or offering unique perspectives or dissenting viewpoints, and for some assessments and EQs, this may be all that is required. As noted in guidance 5.8, findings should not just be based on multiple respondents but also on different types of respondents and different data sources.

Box 6.1: Example paragraph structure

Finding in bold text

- Supporting data points from AMELP data
- Supporting data points from project participant (SSI 15, youth) and donor (SSI 27, USAID)
- Supporting quote from project participant (SSI 15, youth)
- Caveat based on minority, contradicting evidence

6.4 Reports should be organized by findings and not divided by data source. If, for example, a report is based on both a survey and qualitative data, as noted above, the qualitative data should be used to interpret the survey data and vice versa in support of specific findings. There should not be a section for reporting the survey and then a separate section for reporting the qualitative data.

6.5 Expectations about the extent and style of evidentiary support should be discussed early in the evaluation process and not wait until after a draft has been submitted. Recognizing that evaluation users might have diverse expectations about the type of evidentiary support they want to see in a report, expectations should be discussed prior to the work plan/design and again during any outbrief. This is not to say that the study team should always adapt to the users' needs; rather, in some cases, it will be necessary to educate users on why their expectations might not be consistent with the design possibilities. In the early stage of writing, teams should consider submitting as an informal deliverable (i.e., not requiring approval) a short excerpt of a draft section to avoid any mismatch in expectations. This does not need to be extra work for the evaluators, as it is also recommended that team leaders write up an early section of the draft report to model the writing style for their coauthors. These measures can allow for adaptation to any potential report writing concerns at an early stage.

6.6 Take additional steps to proactively build user confidence in the study findings and ensure their utility. There are other ways that study teams can build users' confidence in the findings beyond the pages of the report.

- Instruments presented in work plans and design should include a crosswalk with the original study questions so that users can clearly see how study teams will be answering the questions and conducting the work.
- Two interviewees recommended inviting the USAID core point of contact to participate, when possible, in parts of the team's in-country synthesis/analytical workshop prior to the outbrief.[55] This workshop lasts up to a day and takes place the week of the outbrief. Generally, USAID would attend the last portion of the day once initial analysis is complete. This participation provides an opportunity for the ET to walk key USAID stakeholders through the preliminary findings so that USAID can provide information that the team might have missed and suggest other sources. This also allows the team to begin socializing findings with USAID, so they are not a surprise later.
- The outbrief is also a critical touchpoint with USAID to begin discussing impressions, help the ET understand where USAID's focus and questions are and inform data analysis and writing. In cases where it is premature to provide preliminary findings before more systematic data analysis, it can be helpful to approach the outbrief as a presentation of the issues encountered, initial impressions, analysis approach, and next steps, rather than formal findings. However, it

[55] Interviews with team leaders (3, 16).

is important to align expectations with commissioners in advance about what the outbrief will contain.

- While, in theory, all findings should require the same evidentiary support, recognizing the importance of users' prior assumptions, study teams should be prepared to provide stronger evidentiary support when a finding contradicts prior beliefs. Study teams should seek to better understand users' prior assumptions during early-stage interviews and consultations. There is a risk that this will lead to biasing a study team as evaluators and assessors might feel pressured to deliver findings that confirm users' beliefs; however, these should be considered hypotheses to be transparently tested through the research. Findings should always be driven by the data and never by client preferences.

6.7 Be transparent in the level of confidence in findings. Study teams can use language and visuals to communicate the level of confidence they have in their findings. For example, a finding supported by a desk review, interviews with diverse stakeholders, AMELP data, and survey data is stronger than one based on a handful of similar key informants. This variation in confidence can be expressed both through language and visually. For example, the following phrases from study conclusions show varying levels of confidence:

- "Multiple data sources make clear that…"
- "Based on interview data and the survey responses, the ET can be reasonably confident that…"
- "Some interview data suggests…"
- "Absent more data, the ET can only tentatively conclude that…"
- "There is insufficient evidence to claim that…"

The level of confidence can also be portrayed graphically. For example, Figure 2 provides a color-coding system for the level of confidence in a finding: light green for low confidence, green for moderate confidence, and dark green for high confidence. Figure 3 is similar but allows for a color-coding system of the level of confidence by data source and the possibility that some sources provide contradictory evidence. Each of these data sources presents a column in Figure 3. Imagine that a study relies on secondary literature (symbolized in the figure by the letter L), a coding of past PE reports (C), interviews (I), and a survey (S). In this case, the secondary literature does not contribute to the finding, so L is white. The coding of PEs finds strong evidence to support the findings and is coded as dark green. Interviews provide some moderate support for the finding, so it is coded as light green. In a survey of study commissioners, responses somewhat contradict the finding, so it is coded as light red. Adoption of such an approach should be discussed early in the drafting process, as these visualizations will ultimately be based on the subjective determination of what constitutes moderate or strong evidence.

Figure 2: Color coding of confidence in finding

Finding 1.1	Commissioners often want to know the answers to EQs that cannot be answered effectively through a traditional PE.	

Figure 3: Color coding of confidence in findings by data source

Finding 1.1	Commissioners often want to know the answers to EQs that cannot be answered effectively through a traditional PE.	

GUIDANCE FOR STUDY COMMISSIONERS

1. If a study commissioner has a strong preference for qualitative data presented in a quantified way, then this needs to be discussed early in the design process. If it is not feasible within existing constraints to carry out such an approach (e.g., no representative sample can be drawn, and a highly structured instrument is not desirable), then some flexibility will be required.

2. Read evaluation reports recognizing one's own potential biases. For example, if the reviewer has a prior understanding that a program is effective and expects a report to support that conclusion, this will bias their ability to objectively review the report and to learn from it.

7. CLARITY OF FINDINGS TO FACILITATE USE

SUMMARY

The Challenge: Evaluations, studies, and assessment reports are often lengthy, and key points risk being buried in reports or never read by the intended users. Research teams often struggle to ensure that key points are highlighted without losing needed nuance or adequate empirical support.

Minimal Guidance:

- Provide a summary of the question's response at the outset (7.1).
- Use bolded topic sentences summarizing findings, followed by supporting evidence (7.2).
- Shift much of the methodology explanation section to an annex (7.4)
- Implement processes to ensure a well-written report, including a robust internal review process aided by a checklist (7.6).

THE CHALLENGE

Reports are too long and often not widely read. Both the interviews and PE report review show that the typical qualitative report is long, dense, text heavy, and complex. Figure 4, for example, provides a portion of a zoomed-out report, in which the 30-page finding section did not contain any tables, figures, boxes, or images. Reports often include an overwhelming number of data points that are not well-organized or structured in a way that highlights well-supported findings to readers. Partially, as a result, reports are not read or understood by many of their intended users, undermining the utility of the study. This problem is driven principally by the nature of the task itself. As one interviewee noted, writing about complex things in a simple way is challenging. This is a particular challenge with qualitative research, where it is perhaps more difficult to develop summary visualizations.

Figure 4: Example of text-heavy qualitative report with no figure, table, or image

Source: The Cloudburst Group. (2022). Ukraine DOPP final performance evaluation. USAID.

The challenge is driven as much by users as it is by report authors. On the donor side, the nature of the report is often a product of the questions or scope. USAID guidance is for evaluations to ask between one to five questions,[56] and while many evaluations cap the number of questions at five, questions are often complex and entail a large number of sub-questions. The more questions, the longer the report, and efforts to shorten the report to meet page limits often lead to removing tables, figures, and images that aid readability. In addition, as discussed in Pain Point 6, "Evidentiary support for statements," different study users have different preferences in the kind of data and analysis that they find convincing. For example, some users just want the facts, and others may want more interpretation.

On the research side, several respondents noted that evaluation and assessment report writing require skills and talents that not every study team possesses. One team leader argued that researchers borrow too much from academia and not enough from management consulting. Specifically, the interviewee argued that evaluators should do more to boil down their findings, prioritize among them, draw out the "so what," and present information in an easy-to-understand way.

The dynamic between users and researchers is also an obstacle. One evaluator noted that even though users frequently request an accessible, short report, once this is submitted, they ask for more content. The interviewee went on to note that evaluators throw in the "kitchen sink" so that the donor cannot say that they have left something out. Another team leader seconded the point, noting that reports are long because it is less risky to include everything.

GUIDANCE FOR ETs AND LPs

7.1 Provide a summary of the question's response at the outset. When reports are organized by question, as in the case of evaluation reports, the question should be followed by offset text, such as a text box, that provides a summary of the team's answer to the question. Of course, while it appears first, this text should be written last, as a summary of the findings and conclusions presented in the report section. This text should not contain any content that is not well-substantiated in the text below. This same language can then be used in the executive summary.

7.2 Use bolded topic sentences summarizing findings, followed by supporting evidence. Findings should be presented in bold as a topic sentence to a paragraph. Subsequent text should then provide the data points and evidence to support and explain that finding (See Box 6.1 for an example). This is already a common existing practice that simply needs further scale-up.[57] It avoids problems frequently seen in report writing, including: 1) related data points reappearing at diverse points in the report and 2) various data points organized by topic that do not necessarily support specific findings. This also requires authors to not just present their data or restate what interviewees told them, but also to analyze it and draw specific findings and conclusions from the data. Organizing the narrative under clear findings also requires prioritization and excluding data points that do not necessarily inform a finding or conclusion.

7.3 Use visualizations to summarize qualitative information. Several examples are provided in Table 2, including examples for explaining existing interventions, research methodologies, and qualitative findings. In addition to the quantification of qualitative data, qualitative visualizations can include timelines, Venn diagrams, fishbone diagrams, network maps, word clouds, feedback loop visualizations, flow charts, and journey maps.[58] Tables also offer a way to present qualitative information in a more systematic way than a narrative, and photos offer illustrative potential, including

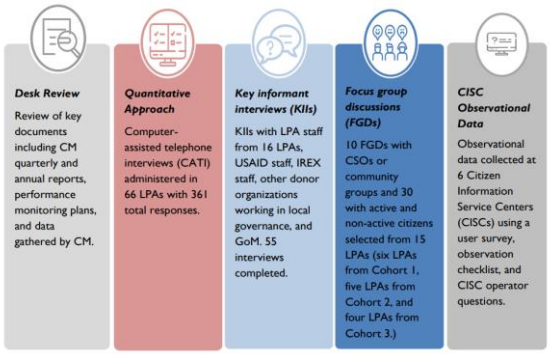
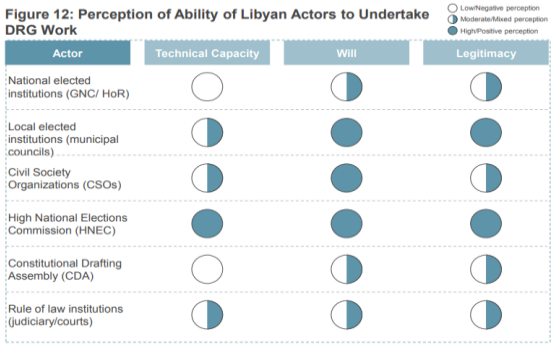
[56] See USAID ADS 201.3.6.7.

[57] The PE report review found that 41 percent (24 or 59) used bolded, italicized findings or otherwise used tools to make key points stand out.

[58] See for example: NVivo. (2021). *Webinar: Creative ways to visualize qualitative data* [Video]. YouTube. https://www.youtube.com/watch?v=MywkhShzAFM; Evergreen, S. (2019). *Effective data visualization: The right chart for the right data*. Sage Publications; Evergreen Data. (n.d.). *Qualitative data visualization*. https://stephanieevergreen.com/qualitative-viz/

before-and-after images. Quotes can also be extracted and made to stand out more from the rest of the text.

Table 2: Options and examples of visualizations

OPTIONS	EXAMPLES
Explain the activity: • Maps of where the activity occurs • Visualizations of the theory of change or logical framework • Tables laying out diverse workstreams	**Figure 1: Comunitatea Mea partner communities by cohort (Comunitatea Mea Annual Report, 2021)** 
Explain the methodology • Summary of data sources • Map of research site locations • Tables of respondent type	**Figure 2: Research design** 
Summarize qualitative findings • Visualizations of subjective assessments • Tables summarizing strengths and weaknesses of activity components • Figures placing actors or activities on a two-by-two plot • Illustrative case studies as text boxes • Visualized quotes • Before-and-after photos	**Figure 12: Perception of Ability of Libyan Actors to Undertake DRG Work** 

Sources: Nayyar-Stone, R., Mark, K., Lupusor, A., Beschieru, I., Wallach, J., Downey, S., Austin, S., & Solovyeva, A.. (2022). Comunitatea mea mid-term performance evaluation. USAID; Wichmann J., Vittum, K., & Kjærum, A. (2016). Mid-term evaluation of democracy, human rights and governance (DRG) programming in Libya and results from a national and urban DRG survey. USAID.

While these visualizations can make a report more attractive and break up the text, the primary goal of visualizations should be to better convey meaning to the reader. As such, this guidance is not suggesting to fill reports with photos for aesthetic appeal, but rather visualizations that aid in conveying key study findings. In most cases, these visualizations do not require additional resources. However, a

graphics designer can be a useful asset for specialized products like infographics or more complicated visualizations.

7.4 Shift much of the methodology explanation to an annex. There is some methodological information that needs to be in the report body for readers to understand the source of the report findings and the report's methodological strengths and weaknesses. As noted in Pain Point 1, "Case or site selection for small-n studies," and Pain Point 2, "Selection of respondents," detailed information about the specifics of site selection, sampling, and measurement—while still important—can be shifted to an annex to shorten the body of reports and maintain a focus on the findings. To be clear, the study team is not arguing to decrease the methodological content overall.

7.5 Conclusions sections should identify implications for decision-making and where actions need to be taken. The conclusion should serve as a bridge between the findings and recommendations, identifying implications for decision-making and where actions need to be taken for the users. In writing the conclusions, authors should return to the purpose or decision-making needs the evaluation was designed to inform. For example, if an evaluation is to inform a future legislative strengthening program, and the current programming is found to be ineffective because of the dominance of the majority party, then the conclusion will note the need to change course in a way that recognizes the majority party dominance. The recommendations will then offer specific and realistic actions to be taken by designated actors. The conclusions should not simply summarize and restate the findings. Under this guidance, such a summary appears at the top of each question and does not need to be repeated.

7.6 Implement processes to ensure a well-written report, including a robust internal review process aided by a checklist. Identify a lead writer and verify that this individual has produced good products in the past. The lead writer does not need to be the team lead. For example, if English is not the team leader's first language, they might need to be paired with a strong English writer to ensure findings are well-communicated.

LPs should employ a robust quality assurance process that includes an internal and substantive "red team" review before drafts are shared with users. This requires that team drafts be finalized one week before submission deadlines. This review should be aided by a checklist of key points, including—for example—the USAID Style Guide recommendations to write at a sixth-grade level and to use active voice. The review should also entail a process of reverse-outlining. This entails pulling out the findings' headings (see 7.2) into a separate document to ensure overall report coherence.

7.7 Develop complementary products that go beyond the report, including targeted briefs, infographics, slides, presentations, videos, or podcasts. Such products should be tailored to specific audiences (e.g., USAID Missions, local government counterparts), aim to facilitate decision-making and the use of study findings, and go beyond the executive summary.[59] For example, this two-pager from a midterm PE of a USAID rapid response mechanism does not attempt to summarize the report. Instead, it provides information and guidance targeted to others in a position to improve how USAID provides rapid response. These products should be envisioned and budgeted at the scoping stage; however, the specifics (e.g., audience, content) should be determined at the report writing or finalization stage.

[59] See also Counterman, M., Conte, S., Starosta, A., Marple-Cantrell, K., Barker, M., & Hatano, R. (2022). *Evidence and learning (E&L) utilization measurement analysis (UMA)*. The Cloudburst Group & USAID.

GUIDANCE FOR STUDY COMMISSIONERS

1. Develop evaluation and study questions to inform decision-making and effectively communicate to study teams about the planned uses of the study.
2. Request complementary products (e.g., targeted briefs) in scopes of work and allocate funds and time for the development of these products.
3. Plan for timelines that include one week for "red team" review and revisions of internal reports. Analysis and drafting require at least four weeks after in-country data collection, but ideally six weeks.
4. Read reports with decision-making and use in mind. While it is easy for reviewers to get bogged down in editorial issues of report structure and writing style, users should remain focused on what the findings mean for their decision-making.

APPENDIX 1: REQUIRED ELEMENTS

WORKPLAN

Site Selection: Clearly articulate the approach, if any, that will be used to select field research sites and ensure its alignment with EQs:

- Why is selection necessary? How many sites are there, and how many will be selected?
- What is the population of sites? What are their main characteristics? Is information about site characteristics available to inform case selection? How was this information obtained?
- What site selection strategy will the ET use and why? How does the selection approach align with the EQs? Is the main goal of the evaluation to learn about how the program is being implemented and experienced "on average" or to learn from specific sites?
- Are there any obvious criteria that should be used to group sites into strata or types as part of site selection? What are the criteria that the ET used?
- Are there any other inputs (e.g., USAID or IP suggestions) that the ET considered to inform case selection?
- Did the ET have to adjust case selection in response to accessibility and security considerations? If they did, what was the process? How were replacement sites selected?

ETs should consider whether their PE could benefit from an incremental or sequenced approach to site selection. If the ET plans to use a sequenced approach, the justification and process to finalize site selection should be included in the work plan/design report.

Respondent Selection: Clearly articulate the approach that will be used to select respondents for individual and group interviews and ensure its alignment with EQs:

- What are the potential types of respondents (e.g., key informants, program participants, and indirect beneficiaries)? Are there any obvious criteria that should be used to group respondents? What are the criteria that the ET used to group respondents?
- For each type of respondent:
    - What is the population? What are their main characteristics? Is information about their characteristics available to inform case selection? How was this information obtained?
    - What is the size of the population, and how many respondents will be selected?
    - What types of potential respondents might be systematically missing from the list, and what bias, if any, could their absence introduce in the findings?
    - What respondent selection strategy will the ET use and why? How does the selection approach align with the EQs? Is the main goal of the evaluation to learn from the "average" respondent or from specific respondents?
- Are there any other inputs (e.g., USAID or IP suggestions) that the ET considered to inform respondent selection?
- What strategies will the ET use to minimize nonresponses? How will the team keep track of nonresponses?

Data Collection Tools:

- Include all data collection consent statements and tools in an annex to the evaluation design/workplan.
- Attach a crosswalk of tool questions/indicators with the original study questions so that users can clearly see how study teams will be answering the EQs.

Data Capture: Clearly articulate the planned systematic method that will be used to capture a close-to-verbatim record of each qualitative event and organize that data:

- Will recordings be used as a backup or primary capture strategy? If recording, will transcriptions be produced? What is the backup data capture strategy in case of refusal to record?
- Who will be responsible for taking notes?
- How will team members be trained to capture notes?
- What quality assurance will be done on notes and by whom?
- How will respondent confidentiality be protected?
- Was ethical and/or research/institutional governance approval obtained for the study? If so, where?
- How was informed consent obtained from participants?
- What language will notes/transcripts be in?
- Will AI tools be used to aid data capture, and if so, how?

Data Analysis: Document-planned systematic process for arriving at findings:

- What types of inquiry and analysis strategy (Box 5.1) will be used?
- Will notes or transcripts be analyzed? How?
- Will data be coded as part of the analysis process? If yes, how? How will coded data be included in the final report?
- Who will perform data coding? Who will perform analysis and interpretation?
- What procedure will be used for teams to confirm findings collectively (among themselves)?
- How will the team establish inter-coder reliability?
- Will AI tools be used to aid data analysis, and if so, how?

Limitations:

- Explain potential limitations or biases stemming from the selection strategies employed (e.g., if excluding low-priority respondents could have substantial implications on findings, missing participants from population lists, etc.).
- Flag potential sources of social desirability bias and outline specific mitigation strategies.

EVALUATION REPORT

EXECUTIVE SUMMARY

Include the summary language from the findings section in the executive summary as described below.

METHODS

Site Selection: Clearly articulate the site selection approach in the main body of the evaluation report and include more detailed information in the annex.

- Note any deviations from original plans (i.e., due to security and accessibility considerations) and the analytical implications of these deviations.
- Note adjustments (e.g., when a large collection of sites or a whole region is inaccessible due to conflict) that needed to happen before site selection that effectively redefined the population of the sites for the PE.

Respondent Selection: Clearly articulate the respondent selection approach in the main body of the evaluation report and include more detailed information in the annex.

- Differentiate between key informants and other respondents.
- Provide selection information for each data collection approach (i.e., interviews, group interviews, and FGDs).
- Define what the population is and how the selected sample relates to the population, especially in the case of program participants and indirect beneficiaries.
- Disclose any prioritization of key informants that was made when selecting respondents.
- Explain any deviations from original respondent selection in response to logistical or security considerations, as well as the implications of these deviations on findings.
- Report response rates and ideally contact and refusal rates, especially for program participants and indirect beneficiaries.

Data Capture and Analysis: Clearly articulate the systematic method used to capture, organize, and analyze data and arrive at findings from all data sources (e.g., desk review, interviews, focus groups, surveys, and secondary data) in the main body of the evaluation report and include more detailed information in the annex.

- Note any later deviations from the data capture plan in the results report methods annex.

LIMITATIONS

See above guidance for the limitations section of the design workplan.

FINDINGS

Implications of Methods:

- Disclose the exclusion of low-priority respondents that could have substantial implications on specific findings.
- Disclose any potential implications of nonresponse in the analysis of specific findings.
- Note any data that is presented that could be subject to social desirability bias.
- Document any significant disagreement findings/interpretation among team members.

Citations:

- Present sufficient data to allow a reader to assess whether or not the interpretation is supported by the data.
- Cite the source of evidence in a way that provides basic information about the source while maintaining confidentiality.

Organization:

- Organize by findings. Do not divide by data source.
- Provide a summary of the question's response at the outset in text that is offset, such as a text box. This text should not contain any content that is not well-substantiated in the text below. This same language can then be used in the executive summary.
- Use bolded topic sentences summarizing findings, followed by supporting evidence.
- Use visualizations to summarize qualitative information.
    - In addition to the quantification of qualitative data, qualitative visualizations can include timelines, Venn diagrams, fishbone diagrams, network maps, word clouds, feedback loop visualizations, flow charts, and journey maps.
    - Tables offer a way to present qualitative information in a more systematic way than a narrative, and photos offer illustrative potential, including before-and-after images.
    - Quotes can be extracted and made to stand out more from the rest of the text.

CONCLUSIONS

Identify implications for decision-making and where actions need to be taken.

- Return to the purpose or decision-making needs the evaluation was designed to inform.
- Offer a suggested course change.
- Do not simply summarize and restate the findings.

ANNEXES

Shift much of the methodology section to an annex to shorten the body of reports and maintain a focus on the findings. As noted above, the methodology annex should include detailed information about the specifics of site selection, sampling, measurement, and analysis.

APPENDIX 2: FULL DESCRIPTION OF STUDY METHODS

The study team followed a sequenced approach, including: 1) identifying the team's initial thoughts about these issues and building on the team's extensive experience; 2) reviewing recent DRG PE reports produced under the LER mechanism; 3) reviewing the academic and practitioner guidance literature; 4) conducting KIIs with evaluation and assessment commissioners, experts at key evaluator and research contracting firms, and past team leaders; and 5) convening workshops to review draft guidance and fill in the remaining gaps. After revising and finalizing the guidance, the team concluded consensus-building efforts with a presentation of the guidance to USAID stakeholders.

Write Out the Team's Initial Thoughts

The principal investigators wrote out their initial thoughts, including descriptions of the seven pain points, minimum standards, and best practices to address them. They presented and discussed their individual thoughts and built consensus around an initial set of team priors. The team also used this discussion to identify areas where additional information was needed and to refine the process to gather such information. This includes the review of PE reports produced under LER mechanisms, the review of specialized qualitative methods literature, and interviews and consultations with various types of informants.

Review of PE Reports

Through a review of recent DRG LER traditional PE reports, the team identified and coded how ETs managed a subset of pain points observable in the PE reports themselves. These include: 1) case selection approach for small-N studies (e.g., selection of sites for visits); 2) selection of interviewees and FGD participants; 3) approach to dealing with SDBs (if any); 4) qualitative data capture strategy; 5) qualitative data analysis strategy; 6) presentation/citation of evidence in findings; and 7) clarity of findings presentation. The team used this information to update their priors as well as to make the case for the need for guidance on minimum standards and best practices.

The review focused on the methodology sections of all PE reports conducted under the LER I and LER II mechanisms. The inventory of PEs is available in Appendix 5 and here. The team used two databases provided by USAID to compile a list of 63 PEs reports: the inventory of Evidence and Learning Output, which is updated through August 2022, and the Learning Harvest Database, which is updated through August 2023. The team accessed the reports using the links in the Development Experience Clearinghouse (DEC) provided in the Learning Harvest Database. For reports not included therein, the team searched the DEC directly. Any reports that were not publicly available were accessed by searching the files of each of the evaluation firms and USAID.

The team reviewed 61 of the 63 PE reports conducted under the LER mechanisms. Two PE reports were excluded because they were not accessible to the team. The team developed a rubric to guide the review process (Appendix 6) and used a Google Form to populate a thematic matrix in Google Sheets to capture information for each reviewed PE report. To ensure consistency, all team members reviewed a randomly selected subset of reports at the beginning of the review process as a training exercise, completed in two rounds. Thereafter, each report was reviewed by only two team members.

The team did not review assessment reports. This is primarily because the nature of case selection and respondent selection is very different between PEs and assessments; however, it is also based on the practical consideration of limited time and budget.

Review of Specialized Literature

The purpose of this review was to inform strategies to address common pain points, minimal standards, and best practices. The team reviewed two types of literature:

1. Social science, especially political science, literature on qualitative methodology
2. Evaluation practitioner literature and resources

The review of academic literature focused on a set of seminal works in the field of qualitative methodology listed in Appendix 7 with the review of some additional publications as deemed necessary. These additional publications were identified using the reference list of seminal works, targeted searches on Google Scholar and online publication catalogs, and recommendations arising from interviews and consultations.

The review of evaluation practitioner literature and resources focused on materials available from donors, think tanks, and CSOs. The team identified a set of websites as the starting point for the search. The preliminary list included the USAID Learning Lab, the Global Evaluation Initiative's BetterEvaluation, the Organisation for Economic Co-operation and Development's resources for the evaluation of development programmes, and the evaluation methods resources from the World Bank's Independent Evaluation Group. Additional materials were identified using the reference list of reviewed documents and recommendations arising from interviews and consultations.

The team used a thematic matrix in Google Sheets to capture relevant information for each reviewed document and facilitate synthesis across pain points and topics of interest.

Interviews and Consultations

The team conducted expert interviews and consultations with thought leaders at the intersection of DRG programming and MEL to: 1) validate and document the common pain points affecting traditional PEs and potentially identify new pain points and 2) aid in identifying and validating minimum guidelines and best practices.

The team selected three categories of respondent to be interviewed: 1) USAID commissioners of evaluations through past and present LER mechanisms and similar mechanisms; 2) the firms implementing these mechanisms and other prominent firms in the DRG evaluation space; and 3) team leaders who had led two or more past LER evaluations or assessments or are particularly well regarded for their methodological expertise. The team began by compiling a list of interviewees fitting these criteria, though it was not exhaustive. Within each category, the team prioritized interviewees according to their expertise and potential to contribute to the utilization of study results. The team conducted approximately 21 interviews with respondents from this list and identified additional respondents via snowball sampling during data collection and through the PE report review. The list of KIIs by category is included in Table A.1, below.

Table A.1: List of interviews by type and gender

INTERVIEW TYPE	NUMBER OF INTERVIEWS	NUMBER OF MEN	NUMBER OF WOMEN
Commissioner	8	5	11
Team Lead	6	5	1
Evaluator/firm	6	3	6
IP MEL Staff	1	0	2

INTERVIEW TYPE	NUMBER OF INTERVIEWS	NUMBER OF MEN	NUMBER OF WOMEN
Total Interviews	**21**	**-**	**-**
Total Interview Participants	**33**	**13**	**20**

Respondents were invited via email to participate in an interview. Interviews took place electronically using Google Meet or Microsoft Teams and, with the consent of the participant, were recorded and automatically transcribed. Recordings and transcriptions will be deleted at the close of the tasking. The team manually coded interviews using a thematic matrix in Google Sheets. The analysis involved reviewing the interview notes and carefully coding and grouping responses in a consistent manner according to similar themes. This allowed the comparison of responses and identification of common trends. Where possible, interviews were attended by all three principal investigators. An interview guide is included in Appendix 8.

Internal Draft Guidance

Upon concluding the review of PE reports, the review of specialized literature, and the interviews and consultations, the team arrived at an internal rough draft guidance document. For each pain point, the document provided preliminary content on: 1) the nature of the challenge; 2) obstacles to addressing the challenges; 3) strategies for overcoming the obstacles; 4) minimum guidance; and 5) best practices. Findings from the review of PE reports were used to provide illustrative examples throughout the document. The document is targeted primarily to the IPs and evaluators conducting traditional PEs and assessments and secondarily to USAID staff commissioning, managing, and using traditional PEs and assessments.

Validation Workshops

After the team synthesized the results of the desk review, expert interviews, and consultations into the draft guidance document, the team conducted a stakeholder validation exercise to validate key elements of the draft guidance, fill in any remaining holes, and start to build consensus around minimum standards and best practices.

This took the form of three 90-minute validation workshops with USAID, IP evaluation experts and past team leaders, and the USAID DRG Learning Community of Practice. All experts identified as potential interviewees were invited to participate in this workshop, as well as a wider group of stakeholders (the USAID DRG Learning Community of Practice) (Table A.2). The workshops occurred virtually and consisted of a presentation of draft guidance and approximately two breakout sessions to solicit initial reactions to the guidance (grouped by pain point) and fill key gaps in a shared Padlet document.

Table A.2 List of workshop participants by gender[60]

VALIDATION WORKSHOP	TOTAL NUMBER OF PARTICIPANTS	NUMBER OF MEN	NUMBER OF WOMEN
1. Evaluation Firms and Team Leads	9	3	6

[60] The breakdown by gender excludes participants for whom gender information was not available.

VALIDATION WORKSHOP	TOTAL NUMBER OF PARTICIPANTS	NUMBER OF MEN	NUMBER OF WOMEN
2. DRG Community of Practice	17	7	9
3. Evaluation Commissioners	8	2	6

Draft Guidance and USAID Dissemination Event

The team revised the internal draft guidance to reflect the learnings from the validation and consensus-building workshops. The product was a draft guidance document for all stakeholders who took part in interviews/consultations and workshops (including the USAID LER II and LER III CORs) for comment. Following a round of comments, the team will lead a focused conversation with the LER II and III CORs and COPs to develop a plan to operationalize the minimal standards and encourage the adoption of best practices in all future LER III taskings.

Final Guidance and Further Dissemination

The team will submit the final guidance document addressing feedback from the LER III CORs and other relevant stakeholders. Upon approval, a 508 compliant version of the document will be uploaded to the DEC. Lastly, to support further dissemination, the team will produce a visually attractive two-page brief and organize an open dissemination event to present the guidance to a wide audience, including USAID and non-USAID stakeholders (e.g., interagency, DRG and LER IPs, and DRG evaluators).

APPENDIX 3: NOTE-TAKING GUIDANCE

The ALEC Note Taking Guidance and Template can be found in the "Appendix 3" folder within this report's zipped file.

APPENDIX 4: RAPID-ANALYSIS CODING MATRIX TEMPLATE

The ALEC Rapid-Analysis Coding Matrix Template can be found in the "Appendix 4" folder within this report's zipped file.

APPENDIX 5: LIST OF LER PEs

The list of PE reports that the team reviewed during the study are listed below.

- Midterm Performance Evaluation of the Liberia Governance and Economic Management Support Program (LER I NORC #5)
- Midterm Evaluation of Counter-Trafficking in Persons II in Cambodia (LER I SI #9)
- Final Performance Evaluation of the Nepal Peace Support Project (LER I NORC #14)
- Midterm Performance Evaluation of the Judicial Strengthening Project in Macedonia (LER I SI #7)
- Midterm Performance Evaluation of USAID/Macedonia's Interethnic Integration in Education Project (LER I SI #11)
- Final Performance Evaluation of the Monitoring Nepal's Peace Process and Constitution Drafting Process Project (LER I NORC #10)
- Final Evaluation of the Land Conflict Resolution Project in Liberia (LER I SI #10)
- Midterm Performance Evaluation of the USAID Media Strengthening Program (LER I NORC #22)
- Performance Evaluation of the Elections and Political Processes Fund (LER I SI #13)
- Midterm Performance Evaluation of the Regional Investigative Journalism Network (LER I SI #14)
- Performance Evaluation of the Consortium for Elections and Political Party Strengthening (CEPPS III LWA) (LER I NORC #19)
- Midterm Evaluation of Democracy, Human Rights, and Governance (DRG) Programming in Libya and Results from a National and Urban DRG Survey (LER I SI #15)
- Performance Evaluation of the Ukrainian Media Project (LER I SI #17)
- LGBTI Global Development Partnership (LER I NORC #65)
- USAID/Center-Run Displaced Children and Orphan's Fund Family Care First Project in Burundi (LER I NORC #3)
- Performance Evaluation of the USAID–CEPPS Syria Pride II Project (LER I NORC #71)
- Bringing Unity, Integrity, and Legitimacy to Democracy in Somalia (LER I SI #23)
- Evaluation of Juvenile Justice Sector Reform Implementation in St. Lucia, St. Kitts and Nevis, and Guyana (LER I SI #19)
- Yetu Initiative Midterm Performance Evaluation (LER I SI #25)
- Strengthening Community Resilience Against Extremism Midterm Performance Evaluation (LER I SI #24)
- USAID/Sierra Leone Women Empowered for Leadership and Development Project Performance Evaluation (LER II NORC #6)
- USAID/Paraguay Democracy and Governance Program Final Performance Evaluation (LER I NORC #2)

- Ukraine Local Governance Project Whole-Of-Project Evaluation (LER I SI #26)
- Strengthening Civil Society Globally Performance Evaluation (LER II NORC #30)
- Accountability in Moldova Whole-of-Project Performance Evaluation (LER I SI #28)
- USAID/Morocco Civil Society Strengthening Program Final Performance Evaluation (LER II NORC #25)
- Malawi Local Government Accountability and Performance Midterm Performance Evaluation (LER II Cloudburst #20)
- Georgia Good Governance Initiative Activity Midterm Performance Evaluation (LER II Cloudburst #25)
- Georgia Advancing CSO Capacities and Engaging Society for Sustainability Project Performance Evaluation (LER II NORC #40)
- Comunitatea Mea Midterm Performance Evaluation (LER II NORC #55)
- Cote d'Ivoire Political Transition and Inclusion Program Final Performance Evaluation (LER II Cloudburst #30)
- Social Movements in Zimbabwe: A Field Assessment and Evaluation (LER II Cloudburst #36)
- Zimbabwe Democracy and Governance Development Objective (DO) Final Performance Evaluation (LER II Cloudburst #41)
- Ukraine Domestic Oversight of Political Processes Final Performance Evaluation (LER II Cloudburst #38)
- Gender-Based Violence (GBV) Portfolio Performance Evaluation: Collective Action to Reduce GBV Final Report (LER II NORC #54)
- Gender-Based Violence Portfolio Performance Evaluation: Better Together Challenge Final Report (LER II NORC #54)
- Performance Evaluation of the USAID Promoting Civic Education and Participation in South Africa Program (LER II NORC 80)
- Gender-Based Violence Portfolio Performance Evaluation: Resilient, Inclusive, and Sustainable Environments Evaluation (LER II NORC #54)
- Cambodia Social Accountability Portfolio Performance Evaluation Report (LER II NORC #43)
- USAID/Peru Transparent Public Investment Integrity Networks Evaluation: Midline Evaluation (LER II NORC #27)
- USAID/Guatemala Community Roots Activity: Final Performance Evaluation (LER II NORC #70)
- Performance Evaluation of USAID's Response to COVID-19–Enabled Corruption (LER II NORC #57)
- Gender-Based Violence Portfolio Performance Evaluation: Women's Economic Empowerment (LER II NORC # 54)
- Mali Justice Project Performance Evaluation (LER II NORC #56)
- Final Performance Evaluation of the USAID Legal Enabling Environment Program II (LER I NORC #7)

- Midterm Performance Evaluation of the USAID-Burundi Youth for Peace-Building Activity (LER I NORC #44)
- USAID/Liberia Elections and Political Transitions, Getting Ready to Lead Activity Assessment (LER I NORC #51)
- Black Sea Trust for Regional Cooperation Performance Evaluation (LER II NORC #4)
- Enhance Non-Governmental Actors and Grassroots Midterm Performance Evaluation (LER II NORC #5)
- Performance Evaluation of the Information Safety and Capacity Project (LER II NORC #12)
- Midterm Evaluation for Sajhedari Bikaas Project (LER I NORC #26)
- Ukraine New Justice Midterm Performance Evaluation (LER II Cloudburst #14)
- Belarus Civil Society Capacity Building Midterm Performance Evaluation (LER II Cloudburst #16)
- Ukraine Democratic Governance East Midterm Performance Evaluation (LER II Cloudburst #21)
- Human Rights Support Mechanism Rapid Response Mechanism Midterm Performance Evaluation (LER II Cloudburst #24)
- Ukraine Health Report Support Midterm Performance Evaluation (LER II Cloudburst #27)
- Media Program in Ukraine Midterm Evaluation (LER II Cloudburst #32)
- Midterm Performance Evaluation of the Global Labor Program (LER I NORC #11)
- Midterm Performance Evaluation of the Information Safety and Capacity Project (LER I NORC #13)
- Strengthening Eurasian News (SEN) Performance Evaluation (LER II NORC #22)

APPENDIX 6: PE REPORT REVIEW RUBRIC

General instructions:

- Please read all instructions and question prompts closely before answering.
- Answering most questions will require you to read a specific section of the report. Do not read other sections of the report to answer.
- It is acceptable to use the search/find tool to answer questions in addition to reading the relevant sections if you find that helpful.
- When answering search/find questions, only search for the terms specified in quotations in the relevant sections.
- Answer all questions based on what you can read and clearly understand from the text. Do not infer the answer from unclear or ambiguous text.
- "Not clear" is a category that is meant to be sparingly used; use only if absolutely necessary.
- If you have any questions or comments about this form, please send an email to the full team.

C1: Small-n (case study selection)

Answer these questions after reading the Methodology, Evaluation Design, or similar section, and, if available, relevant annex focusing on methodology, evaluation design, or similar. Cases/sites/units here refers to the unit for case selection, the main type of entity on which the ET is selecting to make the study more manageable. This generally does not refer to informants, interviewees, respondents, participants, and the like. Often this is a local jurisdiction where the program is implemented.

1. Are cases/sites/units selected from a broader population or does the study examine the full population?

- Full
- Sample
- Not clear

2. What is the case/site/unit of selection?

[Open ended]

3. Is there a justification of why cases were selected?

- Yes
- No
- Not clear

4. Copy and paste language around case/site/unit/sampling.

[Open ended]

5. Does the approach entail generation of a sampling frame or information on the full population from which to sample from? [Note: A sampling frame is a list of the cases/sites/units forming a population from which a sample is taken or cases are selected.]

- Yes
- No
- Not clear

6. Which of the following options best describes the approach followed for case selection?

- Convenience—cases selected are those that were available, accessible, willing to participate, and the like
- Purposive—case selection is driven by intentional criteria that have analytical significance (e.g., locations with high/low levels of ethnic heterogeneity, locations where program was/was not successful)
- Representative—random sample from sampling frame or a collection of cases/sites/units that are typical/representative of the whole population
- Not clear

7. Is the IP credited with playing a role in the actual selection of cases/sites/units? [Note: Answer "no" if the IP's sole role is to provide or help compile a list of all participants/beneficiaries/people.]

- Yes
- No
- Not clear

C2. Beneficiary/respondent selection for KIIs and FGDs

Answer these questions after reading the Methodology, Evaluation Design, or similar section, and, if available, relevant annex focusing on methodology, evaluation design, or similar.

1. Is there a justification for how non-ex officio respondents (people selected because of their role, e.g., COP, COR) were selected? [Note: USAID and IP/Activity staff will generally be included in most PEs. We are interested in the selection of other informants, especially beneficiaries (either institutional or individual) of activity interventions.]

- Yes
- No
- Not clear

2. Copy and paste language about how non-ex officio respondents (people selected because of their role, e.g., COP, COR) were selected.

[Open ended]

3. Does the approach entail generation of a sampling frame or information on the full population from which to sample from? [Note: A sampling frame is a list of the cases/sites/units forming a population from which a sample is taken.]

- Yes
- No
- Not clear

4. Does the approach note an intention to be representative of participants/beneficiaries/people?

- Yes
- No
- Not clear

5. Does the approach note a purposive approach to selection of participants/beneficiaries/people? [Note: Purposive sampling entails selection driven by intentional criteria. Other possible approaches are convenience sampling (whatever works) and representative sample (random sample from sampling frame or a collection of cases/sites/units that are typical/representative of the whole population).]

- Yes

- No
- Not clear

6. Is the IP credited with playing a role in the actual selection of participants/beneficiaries/people? [Note: Answer "no" if the IP's sole role is to provide or help compile a list of all participants/beneficiaries/people.]

- Yes
- No
- Not clear

C3. Dealing with social desirability bias

Answer these questions after reading the Methodology, Evaluation Design, or similar section, and, if available, relevant annex focusing on methodology, evaluation design, or similar.

1. Is there a limitations section to the methodology?

- Yes
- No
- Not clear

2. Is social desirability bias mentioned in the Methodology section or listed as a limitation?

- Yes, by name
- Yes, by concept
- No
- Not clear

3. If social desirability bias is listed/mentioned, is there a mitigation strategy listed?

- Yes
- No
- Not clear

4. Text search: If social desirability is listed, do a text search of "social desirability bias" to see if the concept reappears in the findings, conclusions, or recommendations sections.

- Yes, in more than one place
- Yes, in one place
- No
- Not clear

5. Is there a survey included in the study?

- Yes
- No

6. Skim through the findings section of the report: Does the report feature a figure of survey/interview/FGD results where respondents/informants/participants are asked if the program has been successful/impactful or a similar question?

- Yes
- No

Upload the figure of survey/interview/FGD results where respondents/informants/participants are asked some version of if the program has been successful/impactful.

C4. Qualitative data capture and management

Answer these questions after reading the Methodology, Evaluation Design, or similar section, and, if available, relevant annex focusing on methodology, evaluation design, or similar. Do not answer based on consent forms from instruments included in the annex. Important: These questions refer to qualitative data only.

1. Is there mention of recording interviews/FGDs? [Note: This question does not apply to surveys.]
    - Yes
    - No
    - Not clear
2. Is there mention of transcripts?
    - Yes
    - No
    - Not clear
3. Is there mention of systematic note-taking (e.g., templates, digital, shared files)?
    - Yes
    - No
    - Not clear

C5. Qualitative data analysis

Answer these questions after reading the Methodology, Evaluation Design, or similar section, and, if available, relevant annex focusing on methodology, evaluation design, or similar. Important: These questions refer to qualitative data only.

1. Is there an explanation of how the qualitative data were analyzed?
    - Yes
    - No
    - Not clear
2. How would you classify the way qualitative data was analyzed?
    - Mention of coding but no description
    - Light-touch thematic coding
    - Evidence of rigorous coding (e.g., preset and emergent codes using specialized software)
    - Tally sheets
    - Not clear
    - Other:
3. Copy and paste language about how qualitative data were analyzed.

[Open ended]

C6. Evidentiary support for statements

Answer these questions after reading the EQ2 subsection in the Findings (or Findings and Conclusions) section of the report. Important: These questions refer to interviewees or participants in interviews, group interviews, or FGDs (i.e., to participants in qualitative data collection only), not to survey respondents.

1. Are individual interviews/FGDs cited in some way (e.g., KII 23)? [Note: This refers to whether the authors provide a citation linking specific interviews/FGDs to asserted findings. For example, "Interview evidence suggests that evaluation planning was simply not a high priority, both in Mission-led designs and in designs supported by DRG Center technical staff (KII 1, 11, 13, 14)"]

- Yes
- No
- Not clear

2. Copy/paste example of how individual interviews/FGDs were cited.

[Open ended]

3. Are vague terms like "some" interviewees/participants/etc. or "many" interviewees/participants/etc. used in the report? [Note: You can use word search in addition to reading when answering this question, but make sure you search only the EQ2 subsection.]

- Yes
- No

4. Copy/paste example of how vague terms like "some interviewees" or "many interviewees" used.

[Open ended]

5. Are more specific terms like "a plurality of," "most," "one," "two," or "a minority of" interviewees/participants used in the report? [Note: You can use word search in addition to reading when answering this question, but make sure you search only the EQ2 subsection.]

- Yes
- No

6. Copy/paste example of how more specific terms like "a plurality of interviewees," "most interviewees," "one interviewee," "two," "minority" interviewees were used.

[Open ended]

7. Are very precise terms like "22 percent of interviewees" used in the report?

- Yes
- No

C7. Clarity of findings to facilitate use

Answer these questions after reading the EQ2 subsection in the Findings (or Findings and Conclusions) section of the report.

1. Are findings numbered? [Note: This refers to whether the report uses "Finding 1" or a similar strategy to help findings stand out.]

- Yes
- No
- Not clear

2. Are findings presented as bolded/italicized or topic sentences? [Note: This refers to whether the main findings (or key takeaways or similar) themselves are bolded/italicized or appear as a topic sentence at the top of the section/subsection. Code "no" if the evaluation question, evaluation question excerpt, or a general topic is italicized/bolded.]

- Yes

- No
- Not clear

3. Are there other strategies evident to help findings stand out?

- Yes
- No
- Not clear

4. Copy/paste example of other strategies evident to help findings stand out.

[Open ended]

5. Upload screenshots/images showing other strategies evident to help findings stand out.

APPENDIX 7: BIBLIOGRAPHY

Balbach, E. (1999). *Using case studies to do program evaluation*. California Department of Health Services. https://www.betterevaluation.org/sites/default/files/ProgramEvaluation.pdf

Bamberger, M. (2012). *Introduction to mixed methods in impact evaluation*. [Impact evaluation notes]. Number 3. InterAction & The Rockefeller Foundation. https://www.interaction.org/wp-content/uploads/2019/03/Mixed-Methods-in-Impact-Evaluation-English.pdf

Bamberger, M., & Mabry, L. (2019). *RealWorld evaluation: Working under budget, time, data, and political constraints*. Sage Publications.

Bans-Akutey, A., & Tiimub, B. M. (2021). *Triangulation in research*. Academia Letters. https://doi.org/10.20935/AL3392

Bennett, A., & Elman, C. (2006). Complex causal relations and case study methods: The example of path dependence. *Political Analysis, 14*(3), 250-267. https://doi.org/10.1093/pan/mpj020

Bennett, A., & Elman, C. (2006). Qualitative research: Recent developments in case study methods. *Annual Review of Political Science, 9*, 455-476.

Bergen, N., & Labonté, R. (2020). "Everything is perfect, and we have no problems": Detecting and limiting SDB in qualitative research. *Qualitative Health Research, 30*(5), 783-792.

BetterEvaluation. (n.d.). *Qualitative impact assessment protocol*. https://www.betterevaluation.org/methods-approaches/approaches/qualitative-impact-assessment-protocol

BetterEvaluation. (n.d.). *Thematic coding*. https://www.betterevaluation.org/methods-approaches/methods/thematic-coding

Brady, H. E., & Collier, D. (Eds.). (2010). *Rethinking social inquiry: Diverse tools, shared standards*. Rowman & Littlefield Publishers.

Brinkmann, S., & Kvale, S. (2019). *Doing interviews*. Sage Research Methods. SAGE Publications.

Collier, D., Brady, H. E., & Seawright, J. (2011). A sea change in political methodology. *Newsletter of the American Political Science Association, 9* (1).

Counterman, M., Conte, S., Starosta, A., Marple-Cantrell, K., Barker, M., & Hatano, R. (2022). *Evidence and learning (E&L) utilization measurement analysis (UMA)*. The Cloudburst Group & USAID. https://pdf.usaid.gov/pdf_docs/PA00Z9K1.pdf

Evergreen Data. (n.d.). *Qualitative data visualization*. https://stephanieevergreen.com/qualitative-viz/

Evergreen, S. (2019). *Effective data visualization: The right chart for the right data*. Sage Publications.

Findley, M. G., Starosta, A., & Sabet, D. (2022). *DRG impact evaluation retrospective: Learning from three generations of impact evaluations*. USAID. https://pdf.usaid.gov/pdf_docs/PA00XF3F.pdf

Flick, U. (2018). *Designing qualitative research*. Second Edition. Qualitative Research Kit Series. SAGE Publications.

Gale, R. C., Wu, J., Erhardt, T., Bounthavong, M., Reardon, C. M., Damschroder, L. J., & Midboe, A. M. (2019). Comparison of rapid vs in-depth qualitative analytic methods from a process evaluation of

academic detailing in the Veterans Health Administration. *Implementation Science*, *14*(1), 11. https://doi.org/10.1186/s13012-019-0853-y

Garbarino, S., & Holland, J. (2009). *Quantitative and qualitative methods in impact evaluation and measuring results*. Governance and Social Development Resource Centre. https://www.betterevaluation.org/sites/default/files/EIRS4.pdf

Gibbs, G. R. (2021). *Analyzing qualitative data*. Sage Research Methods. SAGE Publications.

Gilbert, L. S., Jackson, K., & Di Gregorio, S. (2014). Tools for analyzing qualitative data: The history and relevance of qualitative data analysis software. In J. M. Spector, M. D. Merrill, J. Elen, & M. J. Bishop (Eds.), *Handbook of Research on Educational Communications and Technology* (pp. 221-236). Springer. https://doi.org/10.1007/978-1-4614-3185-5_18

Gilardi, F., Alizadeh, M., & Kubli, M. (2023). ChatGPT outperforms crowd-workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, *120*(30). https://doi.org/10.48550/ARXIV.2303.15056

Goertz, G., & Mahoney, J. (2012). *A tale of two cultures: Qualitative and quantitative research in the social sciences.* Princeton University Press.

Gugerty, M. K., Karlan, D., & Welsh, D. (2016*). Goldilocks toolkit: Monitoring for learning and accountability.* Innovations for Poverty Action. https://poverty-action.org/sites/default/files/publications/Goldilocks-Toolkit-Monitoring-for-Learning-and-Accountability_1.pdf

Humphreys, M., & Jacobs, A. M. (2023). *Integrating inferences: Causal models for qualitative and mixed-method research.* Cambridge University Press.

Kapiszewski, D., & Karcher, S. (2021). Transparency in practice in qualitative research. *PS: Political Science & Politics*, *54*(2), 285-291. https://doi.org/10.1017/S1049096520000955

Kusters, C., van Vugt, S., Wigboldus, S., Williams, B., & Woodhill, J. (2017). *Making evaluations matter: A practical guide for evaluators*. Centre for Development Innovation, Wageningen University & Research.

Miles, M. B., & Huberman, A. M. (1994). *Qualitative data analysis: An expanded sourcebook.* Sage Publications.

Morra Imas, L. G. & Rist, R. C. (2009). *The road to results: Designing and conducting effective development evaluations*. World Bank Publications.

Morra, L. G. & A. C. Friedlander. (1999). *Case study evaluations*. Number 2. Operations Evaluation Department (OED) working paper series. World Bank Group. http://documents.worldbank.org/curated/en/323981468753297361/Case-study-evaluations

Nayyar-Stone, R., Mark, K., Lupusor, A., Beschieru, I., Wallach, J., Downey, S., Austin, S., & Solovyeva, A. (2022). *Comunitatea mea mid-term performance evaluation*. USAID.

Neal, J. W., Neal, Z. P., VanDyke, E., & Kornbluh, M. (2015). Expediting the analysis of qualitative data in evaluation: A procedure for the rapid identification of themes from audio recordings (RITA). *American Journal of Evaluation*, *36*(1), 118-132. https://doi.org/10.1177/1098214014536601

NVivo. (2021). *Webinar: Creative ways to visualize qualitative data* [Video]. YouTube. https://www.youtube.com/watch?v=MywkhShzAFM

O'Connor, C., & Joffe, H. (2020). Intercoder reliability in qualitative research: Debates and practical guidelines. *International Journal of Qualitative Methods*, *19*.

Organisation for Economic Co-operation and Development. (2010). *Quality standards for development evaluation.* DAC guidelines and reference series. OECD Publishing.

Organisation for Economic Co-operation and Development. (2021). *Applying evaluation criteria thoughtfully*. OECD Publishing.

Patton, M. Q. (2002). *Qualitative evaluation checklist*. The Evaluation Center. Western Michigan University. https://files.wmich.edu/s3fs-public/attachments/u350/2018/qual-eval-patton.pdf

Ramírez, R., & Brodhead, D. (2013). *Utilization focused evaluation: A primer for evaluators*. Southbound.

Rathje, S., Mirea, D.-M., Sucholutsky, I., Marjieh, R., Robertson, C., and Van Bavel, J. J. (2023). GPT is an effective tool for multilingual psychological text analysis. *Proceedings of the National Academy of Sciences, 121*(34). https://doi.org/10.1073/pnas.2308950121

Seawright, J., & Gerring, J. (2008). Case selection techniques in case study research: A menu of qualitative and quantitative options." *Political Research Quarterly, 61*(2), 294-308.

Sewell, M. (n.d.). *The use of qualitative interviews in evaluation*. CYFERnet–Evaluation, The University of Arizona.

Taylor-Powell, E., & Renner, M. (2003). *Analyzing qualitative data*. (G3658-12). Program Development and Evaluation. University of Wisconsin-Extension.

The Cloudburst Group. (2022). *Ukraine DOPP Final Performance Evaluation*. USAID.

Lewis-Beck, M., Bryman, A., & Futing Liao, T. (2004). Triangulation. In M. Lewis-Beck, A. Bryman, & T. Futing Liao, *The SAGE Encyclopedia of Social Science Research Methods*. Sage Publications, Inc. https://doi.org/10.4135/9781412950589.n1031

USAID. (2013a). *Conducting mixed-method evaluations.* [Technical note]. Version 1. https://2017-2020.usaid.gov/sites/default/files/documents/1870/Mixed_Methods_Evaluations_Technical_Note.pdf

USAID. (2013b). *Evaluative case studies*. [Technical note]. Version 1.0. https://usaidlearninglab.org/system/files/resource/files/usaid_case_study_tech_note_2013.pdf

USAID. (2013c). *Focus group interviews*. [Technical note]. Version 1.0. https://usaidlearninglab.org/system/files/resource/files/focus_group_interviews_tech_note_final_2013_1119.pdf

USAID. (2020). *USAID evaluation policy*. https://www.usaid.gov/sites/default/files/2022-05/Evaluation_Policy_Update_OCT2020_Final.pdf

USAID. (2023). *Assessment of study quality (ASQ) tool*. https://usaidlearninglab.org/system/files/2023-08/usaid_office_of_education_evaluationqualityassessmenttool_dec2017_final.pdf

USAID. (n.d.1). *Toolbox of empowerment: Tools to identify, value, and apply local knowledge*. https://usaidlearninglab.org/system/files/2022-05/cheat_sheets_of_tools_and_methods_for_local_knowledge_0.pdf

USAID. (n.d.2). *C-23: Checklist for use of qualitative methods and approaches*. https://usaidlearninglab.org/system/files/resource/files/mod15_checklist_for_use_of_qualitative_methods_0.pdf

USAID. (n.d.3). *Evaluation toolkit*. USAID Learning Lab. https://usaidlearninglab.org/evaluation/evaluation-toolkit

US General Accounting Office. (1990). *Case Study Evaluations.* https://www.betterevaluation.org/sites/default/files/10_1_9.pdf

Vaessen, J., Lemire, S., & Befani, B. (2020). *Evaluation of international development interventions: An overview of approaches and methods*. Independent Evaluation Group, World Bank. https://documents1.worldbank.org/curated/en/942511608652015232/pdf/Evaluation-of-International-Development-Interventions-An-Overview-of-Approaches-and-Methods.pdf

Vindrola-Padros, Cecilia, and Ginger A. Johnson. 2020. "Rapid Techniques in Qualitative Research: A Critical Review of the Literature." *Qualitative Health Research* 30, no. 10: 1596–1604. https://doi.org/10.1177/1049732320921835

Wichmann J., Vittum, K., & Kjærum, A. (2016). *Mid-term evaluation of democracy, human rights and governance (DRG) programming in Libya and results from a national and urban DRG survey*. USAID. https://pdf.usaid.gov/pdf_docs/pa00kxb2.pdf

APPENDIX 8: INTERVIEW INSTRUMENT

Informed consent: Hello, we are X, Y, and Z with Cloudburst Group, Social Impact, and USAID, and we have asked you for an interview today to inform guidance on how best to address certain challenges in our qualitative evaluation and assessment work. The interview should take approximately an hour, and we will be interviewing or engaging with around 40 other evaluators, assessors, thought leaders, and commissioners of such research. We'll be sharing out the draft and final guidance and report with all interviewees. Interviews are confidential and names of interviewees will not appear in the report or guidance.

- Do you have any questions or any concerns before we begin?
- Do you mind if we record this interview?

Our research is focusing on seven different common challenges. We do not have enough time today to talk about all of these themes, so I would like to prioritize the three that you want to talk about. These might be pain points that you have also wrestled with or have thoughts on how to address. They include:

- C1. Small-n (case study selection)
- C2. Beneficiary/respondent selection for KIIs and FGDs
- C3. Dealing with social desirability bias
- C4. Qualitative data capture and management
- C5. Qualitative data analysis
- C6. Evidentiary support for statements
- C7. Clarity of findings to facilitate use

C1. SMALL-N (CASE STUDY SELECTION)	1. To start, I would love to hear your thoughts on this problem of [small-n case selection]: Is this a significant problem that you have seen in your own work or otherwise? Any examples are particularly welcome.
	2. What does doing [small-n case selection] right look like to you? How would you like to see teams doing this in USAID evaluations and assessments?
	3. What do you see as the obstacles to doing [small-n case selection] right?
	4. Do you see ways to address these obstacles? [This section will include prompts for specific challenges to be refined through the literature review.]
	    a. [prompt] Lack of data/knowledge of the full population
	    b. [prompt] Accessing and using disaggregated IP MEL data
	    c. [prompt] Building consensus around criteria in a purposive sampling approach
	5. Can you point to any exemplary studies (these could be USAID evaluations/assessments but also other sources) where you feel that this has been done well?

C2. BENEFICIARY/RESPONDENT SELECTION FOR KIIs AND FGDs	*Each challenge will include questions 1-5 with some variation in prompts around Q4* 6. Do you see ways to address these obstacles? [This section will include prompts for specific challenges to be refined through the literature review.] a. [prompt] Lack of data/knowledge of the full population b. [prompt] Accessing IP beneficiary data c. [prompt] Absent population information: are there ways to work with IPs to select beneficiaries in a way that doesn't produce biased respondents?
C3. DEALING WITH SOCIAL DESIRABILITY BIAS	7. Do you see ways to address these obstacles? [This section will include prompts for specific challenges to be refined through the literature review.] a. [prompt] Thoughtful question wording b. [prompt] Setting a tone that invites critical responses
C4. QUALITATIVE DATA CAPTURE AND MANAGEMENT	8. Do you see ways to address these obstacles? [This section will include prompts for specific challenges to be refined through the literature review.] a. [prompt] Technology options for transcription and accuracy trade-offs? b. [prompt] When to record, when not to record? c. [prompt] Team leaders/members resistance (e.g., used to pen and paper)
C5. QUALITATIVE DATA ANALYSIS	9. Do you see ways to address these obstacles? [This section will include prompts for specific challenges to be refined through the literature review.] a. [prompt] When do the benefits of coding outweigh the costs? b. [prompt] Time saving (e.g., junior staff coding) vs. quality trade-offs c. [prompt] Team leaders/member resistance
C6. EVIDENTIARY SUPPORT FOR STATEMENTS	10. Do you see ways to address these obstacles? [This section will include prompts for specific challenges to be refined through the literature review.] a. [prompt] How do you convey the level of confidence in a finding without saying X of Y respondents? b. [prompt] Precision vs. "many," "some," "few" c. [prompt] Are quotes evidence or illustrative of an already supported point?

C7. CLARITY OF FINDINGS TO FACILITATE USE	11. Do you see ways to address these obstacles? [This section will include prompts for specific challenges to be refined through the literature review.] a. [prompt] How do you convey the level of confidence in a finding? b. [prompt] Standardization vs. flexibility c. [prompt] Clarity vs. nuance

UNITED STATES AGENCY FOR INTERNATIONAL DEVELOPMENT
1300 PENNSYLVANIA AVENUE, NW
WASHINGTON, DC 20523