



USAID
FROM THE AMERICAN PEOPLE



META-EVALUATION OF QUALITY AND COVERAGE OF USAID EVALUATIONS 2009 – 2012

August 2013

This publication was produced for review by the United States Agency for International Development. It was prepared by Molly Hageboeck, Micah Frumkin and Stephanie Monschein, Management Systems International under subcontract to DevTech Systems, Inc. for USAID Contract No. AID-OAA-M-11-00026.

Meta-Evaluation of Quality and Coverage of USAID Evaluations

2009-2012

Prepared by:

Molly Hageboeck, Micah Frumkin, and Stephanie Monschein, Management Systems International

Contracted under AID-OAA-M-11-00026

Program Cycle Service Center

Prime Contractor: DevTech Systems, Inc.

DISCLAIMER

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

ACKNOWLEDGEMENTS

A study of this magnitude could not have been successfully completed without the hard work and dedication of a great many people. The team would like to thank USAID/PPL/LER, and specifically Cynthia Clapp-Wincek, Rozanne Larsen, and MiKell Brough-Stevenson, for deciding to undertake such a study and for their continuous support and guidance throughout the process. Outside of USAID, Al Smith at the Program Cycle Service Center provided valuable substantive and procedural support. USAID's Development Experience Clearing House (DEC) staff provided us with invaluable assistance in locating evaluations and deserves credit for the improvements it has made in this collection over the past few years.

MSI would also like to acknowledge our expert and incredibly hard working team of researchers and coders without whom we never would have been able to achieve the ambitious task set before us. Among them we include Adam Peterson, Garrison Spencer, Gwynne Zodrow, Ingrid Orvedal, Jeremy Gans, Leah Sly, Liz Freudenberg, Mary Beth Yarbrough, Paul Diegert, Sarah Fuller, Arielle Snyder, and Clare Carlo. In addition to these individuals, the team for this meta-evaluation included Micah Frumkin, team manager and senior coder; Stephanie Monschein, quantitative data analyst for the study; and Molly Hageboeck, the study's Technical Director.

Lastly, we would like to thank the USAID staff from Regional and Technical Bureaus and representatives from 16 firms that conduct evaluations for USAID who took the time out of their busy schedules to meet with us and provide us with invaluable comments and perspectives which helped us to fully understand the state of the quality of evaluations at USAID. We also thank the 25 evaluation Team Leaders who responded to a survey we fielded to obtain information about their recent USAID evaluation experiences and perceptions of changes in the USAID evaluation process.

TABLE OF CONTENTS

Acknowledgements	i
Acronym List.....	vii
Executive Summary	viii
PART I – META-EVALUATION: MAIN FINDINGS	1
1. Introduction.....	2
Background and Purpose	2
Meta Evaluation Questions	2
2. Findings.....	4
A. Evaluation Characteristics.....	4
B. Evaluation Quality Ratings	10
Question 1. To what degree have quality aspects of USAID’s evaluation reports, and underlying practices, changed over time?.....	11
Question 2. At this point in time, on which evaluation quality aspects or factors do USAID’s evaluation reports excel and where are they falling short?.....	20
Question 3. What can be determined about the overall quality of USAID evaluation reports and where do the greatest opportunities for improvement lie?.....	24
3. Conclusions.....	27
4. Recommendations.....	33
PART 2. DETAILED META-EVALUATION FINDINGS	38
1. Basic Evaluation Characteristics in Depth	39
A. USAID Region or Bureau	40
B. Sector or Topic of the Program or Project Evaluated.....	43
C. Scope or Scale of the Evaluation.....	43
D. Evaluation Timing.....	44
F. Type of Evaluation.....	46
I. Evaluation Cost.....	52
J. Duration of the Evaluation	54
2. Evaluation Quality Ratings in Depth	55
Question 1: To what degree have quality aspects of USAID’s evaluation reports, and underlying practices, changed over time?	55
A. Executive Summary Accurately Summarizes Critical Elements of Report Ratings....	58
B. Project/Program Background Ratings	60
C. Description of the Project or Program’s “Theory of Change” Ratings	61
D. Evaluation Purpose Ratings	61
E. Evaluation Questions Ratings.....	62
F. Team Composition Ratings	65
G. Team Awareness of USAID Evaluation Standards Ratings	70
H. Data Collection and Analysis Methods Ratings.....	71
I. Study Limitations Ratings	79

J. Findings Ratings.....	80
K. Recommendations Ratings.....	90
L. Annexes Ratings.....	93
M. Summary Tables on Quality Factors	97
Question 2: At this point in time, on which evaluation quality aspects or factors do USAID's evaluation reports excel and where are they falling short?	104
Question 3: What can be determined about the overall quality of USAID evaluation reports and where are the greatest opportunities for improvement?	115
Annexes	125
Annex A. Meta-Evaluation Statement of Work.....	125
Annex B. Methods	126
Annex C. Meta-Evaluation Rating Instruments	140
Annex D. Evaluation Team Leader Perception Survey Responses	173
Annex E. Group Interview Transcript Summaries.....	188
Annex F. Team Composition	195
Annex G. Bibliography	197

TABLE OF TABLES

Table 1. Chronology of USAID Meta-Evaluations	3
Table 2. Number of Questions Asked in Evaluations	9
Table 3. Quality Factors With the Most Improvement Between 2009 and 2012	11
Table 4. Factors With More Than a 1 Percent Decline in Quality	12
Table 5. Evaluation Quality Ratings for all Factors	12
Table 6. Team Members Identified in Evaluation Reports	15
Table 7. Data Collection Methods Used in 2009–12 Evaluations	16
Table 8. Performance on Evaluation Factor Quality Ratings by Cluster	20
Table 9. Quality Factor Ratings for 2012	20
Table 10. Degree of Association Between Evaluation Scores and Other Evaluation Characteristics	25
Table 11. Recapitulation of Quality Factors That USAID Determines	28
Table 12. Recapitulation of Quality Factors Determined Jointly by USAID and Teams	30
Table 13. Recapitulation of Quality Factors Determined by Evaluation Teams	31
Table 14. Historical Data on Evaluation Purposes	46
Table 15. Sample Cost Section of USAID Form AID 1330-5 (10/87) Prepared in 1995	52
Table 16. Cost of USAID Evaluations in 1987–1988	53
Table 17. Net Change in Evaluation Quality Factor Ratings Between 2009 and 2012	57
Table 18. Historical Presence of Executive Summaries	59
Table 19. Executive Summary Accurately Reflects Report	60
Table 20. Inclusion of Program/Project Background in Evaluation Reports	61
Table 21. Inclusion of "Theory of Change" in Evaluation Reports	61
Table 22. Management Purpose Identified	62
Table 23. Evaluation Questions Addressed Were Identical to the SOW	64
Table 24. Evaluation Questions Were Linked to a Management Purpose	65
Table 25. External Team Leaders	66
Table 26. Historical Meta-Evaluation	66
Table 27. Evaluation Specialists on Teams	67
Table 28. Reports That Identified an Evaluation Specialist on the Team, By Region	67
Table 29. Local Team Members Involved	68
Table 30. Evaluation Reports That Identified Local Team Members, By Region	68
Table 31. 2011–12 Evaluation SOWs That Included Evaluation Policy Appendix 1	70
Table 32. Data Collection Methods Described in Evaluations	72
Table 33. "Getting to Answers" Matrix	73
Table 34. Evaluation Explained Linkages Between Data Collection Methods and Evaluation Questions	73
Table 35. Planned and Actual Use of Evaluation Data Collection Methods	74
Table 36. Evaluation Methods Identified in Previous USAID Meta-Evaluations	75
Table 37. Data Analysis Methods Described in Evaluations, By Year	77
Table 38. Evaluation Explained Linkages Between Data Analysis Methods and Evaluation Questions	77
Table 39. Planned and Actual Use of Evaluation Data Collection Methods	78
Table 40. Reports Includes a Description of Study Limitations	79
Table 41. Report Presented Findings in Relation to Evaluation Questions	81
Table 42. Evaluation Questions were Addressed in the Body of the Report	81
Table 43. Findings Appear to Reflect the Use of Social Science Methods	82
Table 44. Findings Clearly Drew on the Full Range of Data Collection Methods Used	83
Table 45. Quantitative Data Reported as Precise Numbers (Not as "Some," "Many," or "Most")	83

Table 46. Evaluation Findings Were Distinguished From Conclusions and Recommendations.....	84
Table 47. Findings Were Disaggregated By Sex At All Levels	85
Table 48. ADS Guidance on Sex Disaggregation in Evaluations Over Time.....	86
Table 49. Evaluation Questions Addressed Differential Access/Benefits by Gender	87
Table 50. Evaluation Addressed Unplanned/Unanticipated Results.....	88
Table 51. Questions Addressed Alternative Possible Causes of Observed Results.....	89
Table 52. Recommendations Stood Alone in Report.....	90
Table 53. Recommendations Were Clearly Supported by Evaluation Findings.....	91
Table 54. Recommendations Were Specific About Actions To Be Taken.....	92
Table 55. Recommendations Were Clearly Directed to Specific Parties.....	92
Table 56. Evaluation SOW Was Included as a Report Annex.....	93
Table 57. Historical Data on Evaluation Statements of Work (SOWs).....	93
Table 58. Annex Included a List of Sources.....	94
Table 59. Annex Included Evaluation Data Collection Instruments.....	95
Table 60. Report Included Conflict of Interest Forms or Indicated That They Were Available	96
Table 61. Report Indicated How Data Obtained by the Evaluation Will Be Transferred to USAID	96
Table 62. Evaluation Included a Statement of Differences as an Annex.....	97
Table 63. Quality Factor Ratings by Region.....	98
Table 64. Quality Factor Ratings By Sector	100
Table 65. Presence of USAID Forward Evaluations By Region, July 2011 Through December 2012	102
Table 66. Ratings of USAID Forward Evaluations Compared With Non-USAID Forward Evaluations	102
Table 67. Evaluation Quality Factor Rating Clusters	105
Table 68. Quality Factors Clustered into Four Performance Levels.....	106
Table 69. Comparison of Average Evaluation Scores by Bureau, 1983 and 2009-12	117
Table 70. Average Scores by Sector	118
Table 71. Average Scores for USAID Forward and Other Evaluations	119
Table 72. Correlation Between Scores and Other Evaluation Characteristics.....	119
Table 73. The Effect of Evaluation Specialists and Local Team Members on Evaluation Scores	120
Table 74. Quality Factors Associated with the Presence of an Evaluation Specialist	121
Table 75. Quality Factors Associated with the Presence of Local Evaluation Team Members	121
Table 76. Quality Factors Associated with the Presence of an External Team Leader	122
Table 77. Average Scores by Numbers of Questions	123
Table 78. Quality Factors in the Overall Score, Ranked by Degree of Association with the Overall Score	123

TABLE OF FIGURES

Figure 1. Timeline of Evaluation-Related Events at USAID 2009–12.....	2
Figure 2. Responsibilities for Evaluation Quality Factors.....	3
Figure 3. Trends in Number of Evaluations.....	5
Figure 4. Geographic Distribution of 2009–12 Evaluations.....	6
Figure 5. Sector Representation in the Meta-Evaluation Sample.....	6
Figure 6. Non-Experimental Methods Can Help Identify Other Causes.....	8
Figure 7. Evaluation Purposes.....	8
Figure 8. Evaluation Reports that Included Evaluation Questions.....	9
Figure 9. Percentage of Evaluations That Included a Statement of Study Limitations.....	17
Figure 10. Percentage of Evaluations with Specific Overall Scores where an Evaluation Specialist was a Team Member.....	26
Figure 11. Distribution of Evaluations by Region.....	40
Figure 12. AfPak Region Evaluations.....	41
Figure 13. Africa Region Evaluations.....	41
Figure 14. Asia Region Evaluations.....	41
Figure 15. E&E Region Evaluations.....	42
Figure 16. LAC Region Evaluations in the Sample.....	42
Figure 17. ME Region Evaluation in the Sample.....	42
Figure 18. Distribution of Evaluations by Sector.....	43
Figure 19. Scope of USAID Evaluations.....	44
Figure 20. Evaluations With a Scope Other than "Single Project".....	44
Figure 21. Timing of USAID Evaluations.....	45
Figure 22. Percentage By Type.....	46
Figure 23. Percentage of Evaluation Questions, By Numbers of Questions.....	49
Figure 24. Number of Evaluation Questions, By Year.....	50
Figure 25. Numbers of Questions in SOWs for.....	50
Figure 26. Number of Questions, in Clusters, By Region.....	51
Figure 27. Number of Questions, in Clusters, By Sector.....	51
Figure 28. Time for Recent Evaluations.....	54
Figure 29. Resources for Recent Evaluations.....	54
Figure 30. Distribution of What Evaluations Were Asked to Address.....	63
Figure 31. Evaluations that Included Questions for Evaluation Teams.....	63
Figure 32. Presence of Evaluation Specialists and Local.....	69
Figure 33. Factors That Affect Selection of Evaluation Team Leaders.....	69
Figure 34. Information on USAID Evaluation Standards Received By Recent USAID Evaluation Team Leaders.....	71
Figure 35. Presentation of Data Collection &.....	72
Figure 36. Roles in Selecting Data.....	76
Figure 37. Time Available to Prepare for Data Collection Before Field Work.....	76
Figure 38. Time Available for Data Collection and Analysis – Recent Evaluation Experience.....	79
Figure 39. Quality Clusters by Region.....	109
Figure 40. Quality Clusters by Sector.....	109
Figure 41. Quality Clusters For USAID Forward and Non-USAID Forward Evaluations.....	110
Figure 42. Quality of Recent Reviews of Draft Evaluation Reports.....	113
Figure 43. Percentage of Evaluations Designated as USAID Forward Evaluations.....	114
Figure 44. Distribution of Evaluations by Score.....	116
Figure 45. Trends in Overall Scores, 2009-12.....	117
Figure 46. Regional Scores Compared to the Overall Score.....	118
Figure 47. Sector Average Evaluation Scores Compared to the Overall Average Score.....	118

ACRONYM LIST

ADS	Automated Directives System
AfPak	Afghanistan and Pakistan
CNP	Conditions Not Present
AFR	Africa
DEC	Development Experience Clearinghouse
DC	Washington, D.C.
DG	Democracy and Governance
E&E	Eastern Europe and Eurasia
EG	Economic Growth
IRR	Inter-Rater Reliability
LAC	Latin America & Caribbean
M&E	Monitoring and Evaluation
ME	Middle East
MSI	Management Systems International
OAA	USAID's Office of Acquisition and Assistance
OECD/DAC	Development Assistance Committee of the Organization for Economic Co-operation and Development
PPL/LER	Policy Planning & Learning/Learning, Evaluation, & Research
SOW	Statement of Work
UNICEF	United Nations Children's Fund
USAID	United States Agency for International Development
NGO	Non-Governmental Organization
RFP	Request for Proposal

EXECUTIVE SUMMARY

Context and Purpose

This evaluation of evaluations, or meta-evaluation, was undertaken to assess the quality of USAID's evaluation reports. The study builds on USAID's practice of periodically examining evaluation quality to identify opportunities for improvement. It covers USAID evaluations completed between January 2009 and December 2012. During this four-year period, USAID launched an ambitious effort called USAID Forward, which aims to integrate all aspects of the Agency's programming approach, including program and project evaluations, into a modern, evidence-based system for realizing development results. A key element of this initiative is USAID's Evaluation Policy, released in January 2011.

Meta-Evaluation Questions

The meta-evaluation on which this volume reports systematically examined 340 randomly selected evaluations and gathered qualitative data from USAID staff and evaluators to address three questions:

1. To what degree have quality aspects of USAID's evaluation reports, and underlying practices, changed over time?
2. At this point in time, on which evaluation quality aspects or factors do USAID's evaluation reports excel and where are they falling short?
3. What can be determined about the overall quality of USAID evaluation reports and where do the greatest opportunities for improvement lie?

Meta-Evaluation Methodology and Study Limitations

The framework for this study recognizes that undertaking an evaluation involves a partnership between the client for an evaluation (USAID) and the evaluation team. Each party plays an important role in ensuring overall quality. Information on basic characteristics and quality aspects of 340 randomly selected USAID evaluation reports was a primary source for this study. Quality aspects of these evaluations were assessed using a 37-element checklist. Conclusions reached by the meta-evaluation also drew from results of four small-group interviews with staff from USAID's technical and regional bureaus in Washington, 15 organizations that carry out evaluations for USAID, and a survey of 25 team leaders of recent USAID evaluations. MSI used chi-square and t-tests to analyze rating data. Qualitative data were analyzed using content analyses. No specific study limitation unduly hampered MSI's ability to obtain or analyze data needed to address the three meta-evaluation questions. Nonetheless, the study would have benefited from reliable data on the cost and duration of evaluations, survey or conference-call interviews with USAID Mission staff, and the consistent inclusion of the names of evaluation team leaders in evaluation reports.

Characteristics of Evaluations in the Meta-Evaluation Sample

The study sample represents every geographic region and technical area where USAID works. The largest segment of the evaluations (38 percent) was conducted in Africa. On the technical side, the largest segment (29 percent) was health program and project evaluations. The dominance of these evaluations in the study sample is consistent with USAID's allocation of development assistance funds.

Over the years covered by this meta-evaluation, the number of evaluations USAID completed annually rose each year, after a decade of decline that culminated in an all-time low of 73 evaluations in 2007. By 2012, the final year examined in this meta-evaluation, the number of evaluations completed rose to 201.

With respect to the scope of evaluations, the largest portion (76 percent) of evaluations focused on single projects. By type, 97 percent were performance evaluations and 3 percent were impact evaluations that included a comparison group to help determine what would have occurred in the absence of USAID's assistance. This distribution is consistent with USAID expectations, given that impact evaluations of this type are new to USAID and often take several years to complete. Among performance evaluations, there was a roughly equal split between evaluations conducted during implementation and those carried out toward the end of a project or program.

Evaluation Quality Findings

Findings from the meta-evaluation are organized to answer the three questions this study addressed.

Question 1. To what degree have quality aspects of USAID's evaluation reports, and underlying practices, changed over time?

Over the four years covered by the meta-evaluation, there were clear improvements in the quality of USAID evaluation reports. On 25 of 37 (68 percent) evaluation quality factors rated, evaluations completed in 2012 showed a positive net increase over 2009 evaluations in the number that met USAID quality standards on those factors. Ratings on several factors improved by more than 10 percentage points, including *whether findings were supported by data from a range of methods*, *study limitations were identified*, and *clear distinctions were made between findings, conclusions, and recommendations*. Improvements in evaluation quality factor ratings did not generally rise in a linear fashion, but instead fluctuated from year to year. Not all evaluation rating quality factors improved over the study period. MSI, in addition to examining changes over time for the study sample as a whole, assessed changes between 2009 and 2012 on a regional basis, by sector, and for a subset of USAID Forward evaluations to which the Agency, after July 2011, paid special attention from a quality perspective. A t-test was used to compare USAID Forward evaluations with other evaluations. Its results were not significant.

Question 2. At this point in time, on which evaluation quality aspects or factors do USAID's evaluation reports excel and where are they falling short?

Four clusters of evaluation ratings were used to determine where USAID excels on evaluation quality and where improvements are warranted. Evaluation quality factors on which 80 percent or more USAID evaluations met USAID standards were coded as "good." Of 37 evaluation quality factors examined, 24 percent merited the status designation "good." Quality standards for which 50 percent to 79 percent of evaluations were rated positively were designated as "fair." USAID performance was either "good" or "fair" on half of the factors rated. On the remaining evaluation quality factors, USAID performance was deemed "marginal" on 20 percent of those factors and "weak" on 32 percent. Among evaluation quality factors on which compliance was "weak," MSI found that half addressed quality standards that had recently been introduced in USAID Evaluation Policy. Performance on these factors is likely to improve as familiarity with these new standards improves. Among factors rated weak, the most significant involve low levels of compliance with USAID's requirement for the participation of an evaluation specialist on every evaluation team and its expectation that, wherever relevant, data on the results of USAID evaluations will be documented on a sex-disaggregated basis.

Question 3. What can be determined about the overall quality of USAID evaluation reports and where do the greatest opportunities for improvement lie?

On an overall evaluation quality “score” based on 11 of the meta-evaluation’s quality rating factors, USAID evaluations averaged 5.93 on a 10-point scale—with a mode of 7 points and a relatively normal distribution. Statistical tests conducted using this overall score showed that USAID evaluations completed in 2012 were of significantly higher quality than those completed in 2009. MSI also found that evaluations reporting an evaluation specialist as a team member had higher overall quality scores than evaluations where an evaluation specialist was not reported to be involved. This finding was statistically significant at .05, .01, and .001 levels. Other comparisons were not found to be statistically significant.

Conclusions

The overall picture of evaluation quality at USAID from this study is one of improvement over the study period, with strong gains emerging on key factors between 2010 and 2012. The number of evaluations per year increased, and the quality of evaluation reports has improved. While this portrait is largely positive, the study also identified evaluation quality factors, or standards, that USAID evaluation reports do not yet meet. On several core evaluation quality standards—such as clear distinctions among evaluation findings, conclusions, and recommendations—performance was found to be below USAID standards. Other significant deficiencies included the small percentage of evaluations that indicated that an evaluation specialist was a member of the evaluation team, which USAID has required for the better part of a decade, and low ratings on the presence of sex-disaggregated data at all results levels—not simply for input level activities. Low ratings were also found for several evaluation standards introduced in the 2011 Evaluation Policy, but this may simply reflect slow uptake or lack of awareness of standards.

Recommendations

USAID’s broad evaluation improvement initiative already focuses on policies, procedures, guidelines, and training that are intended to change USAID staff knowledge and practices, and through them the practices of organizations and individuals that undertake evaluations for the Agency. Recommendations from this meta-evaluation, to be helpful, must supplement rather than duplicate those efforts. With this in mind, MSI offers three recommendations to USAID/PPL/LER that can significantly enhance the quality of USAID evaluation reports in those areas that offer opportunities for improvement:

- **Recommendation 1.** Increase the percentage of USAID evaluations that have an evaluation specialist as a fulltime team member with defined responsibilities for ensuring that USAID evaluation report standards are met from roughly 20 percent as of 2012 to 80 percent or more.
- **Recommendation 2.** Intervene with appropriate guidance, tools, and self-training materials to dramatically increase the effectiveness of existing USAID evaluation management and quality control processes.
- **Recommendation 3.** As a special effort, in collaboration with USAID’s Office of Gender Equality and Women’s Empowerment, invest in the development of practitioner guidance materials specific to evaluation.

Of these three recommendations, the first is considered the most important for systematically raising the quality of evaluations across all sectors and regions. MSI’s second recommendation is intended to complement its first recommendation and encourage USAID to scale up evaluation management “good practices” already known within the Agency.

PART I – META-EVALUATION: MAIN FINDINGS

This meta-evaluation report provides the United States Agency for International Development (USAID) with a rich array of data to call on as it continues its efforts to strengthen evaluation quality at the Agency. At the same time, MSI recognizes that USAID managers need access to a succinct presentation of the meta-evaluation’s findings. Accordingly, this report is divided into two distinct parts.

Part I introduces the meta-evaluation, summarizes the characteristics of the 340 USAID evaluations conducted between 2009 and 2012 and analyzed by the meta-evaluation team, and answers three questions about the quality of these evaluations. Conclusions and the meta-evaluation’s recommendations for USAID are presented at the end of Part I.

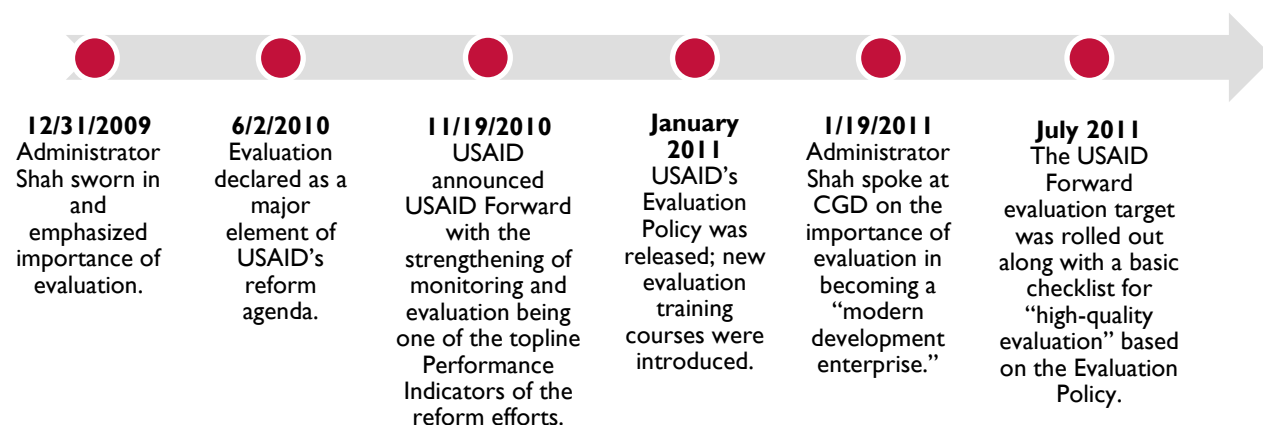
Part 2 provides detailed findings from the meta-evaluation, structured not only on an overall basis, but also on an annual basis. Additional findings are also provided that look at the data through regional and sectoral perspectives. Part 2 also includes evaluations designated by the Agency as USAID Forward evaluations, for which special efforts were made to ensure high-quality products in Missions worldwide and how they compare with other USAID evaluations completed during the final 18 months of the meta-evaluation period.

I. INTRODUCTION

Background and Purpose

USAID views evaluation as playing a critical role in the program cycle, providing evidence to support program and project design decisions, and guiding the implementation of ongoing activities. To these ends, since 2010, USAID Administrator Dr. Rajiv Shah and Agency staff have invested in a range of activities aimed at improving the quality of USAID evaluations, and thus their usefulness. Figure 1 highlights these investments, including the issuance of USAID’s Evaluation Policy in January 2011, the development of new evaluation courses that provided 1,200 USAID staff members and other stakeholders with professional training in evaluation between February 2011 and July 2013, and increased attention on high-quality evaluations through the Agency’s ongoing USAID Forward initiative.

Figure 1. Timeline of Evaluation-Related Events at USAID 2009–12



This evaluation of evaluations, or meta-evaluation, was undertaken to assess the status of USAID’s evaluation practice. The study builds on USAID’s 30-year-old practice of periodically examining evaluation quality to identify opportunities for improvement. Table 1 summarizes that history.

Meta Evaluation Questions

The meta-evaluation discussed in this study covers evaluations completed between January 2009 and December 2012, as discussed in the study’s Statement of Work (SOW) in Annex A. In addition to profiling basic characteristics of evaluations during this period, it addresses three specific questions about their quality:

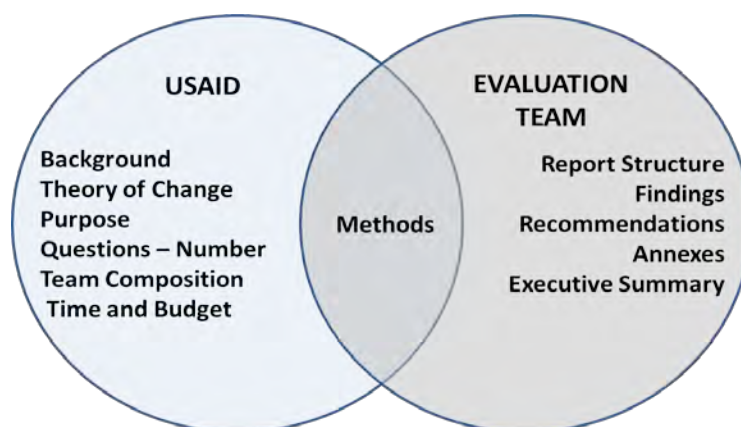
1. To what degree have quality aspects of USAID’s evaluation reports, and underlying practices, changed over time?
2. At this point in time, on which evaluation quality aspects or factors do USAID’s evaluation reports excel and where are they falling short?
3. What can be determined about the overall quality of USAID evaluation reports and where do the greatest opportunities for improvement lie?

Table 1. Chronology of USAID Meta-Evaluations

Previous USAID Meta-Evaluations	
Year	Authors
1982	Triton Corporation
1983	Triton Corporation
1987–88	Development Associates
1989–90	Management Systems International (MSI)
1993–97	Greene
1998–99	Clapp–Wincek and Blue
2005–08	Management Systems International (MSI)
2009 only	Kumar and Eriksson
2009–12	Management Systems International (MSI)

Study Approach

Evaluation quality is a multifaceted concept. It encompasses the methods used to assemble credible evidence to address questions that go beyond what USAID can learn from monitoring program and project performance against predetermined targets, but it does not end there. Quality also involves the transformation of findings, through a clear and transparent reasoning process, into actionable recommendations. Undertaking an evaluation involves a partnership between the client for an evaluation and evaluation team, as Figure 2 suggests.

Figure 2. Responsibilities for Evaluation Quality Factors

Information on basic characteristics and quality aspects of 340 randomly selected USAID evaluation reports was a primary source of evidence for this study. Quality aspects of these evaluations were assessed using a 37-element checklist. Conclusions reached by the meta-evaluation also drew on the results of small group interviews with staff from USAID’s technical and regional bureaus in Washington, and with organizations that carry out evaluations for USAID, as well as from a survey of team leaders of recent USAID evaluations. A full description of the methodology for this study, including study limitations, is provided in Annex B.

2. FINDINGS

In parallel with Figure 2 above, USAID evaluation characteristics discussed in section A below are largely determined by USAID staff. Quality aspects presented in section B reflect decisions made primarily by evaluation teams. Also in section B are methods, to which both parties contribute.

A. Evaluation Characteristics

USAID staff decisions define most of the basic characteristics of USAID's evaluation portfolio. Such characteristics include the programs and projects being evaluated, the scope of those evaluations, their timing, the type of evaluation to be conducted, the number of questions to be addressed, the composition of the evaluation team, initial ideas about evaluation methods, and the evaluation cost and duration. These characteristics for USAID evaluations completed between 2009 and 2012 are described below.

Number and Distribution of Evaluations

Increasing the number of evaluations USAID conducts in a year was not an explicit aim of the Agency's evaluation quality improvement initiative, but nevertheless it occurred. From 2009 to 2012, the number of USAID evaluations nearly doubled from 112 in 2009 to 201 in 2012.^{*} This is after the number began to recover in 2008 following a 12-year decline, depicted in the red line in Figure 3.[†] The most significant gains in recent years emerged between 2010 and 2012.

In this same figure, a blue line displays USAID's historical record of documents coded as evaluations in USAID's Development Experience Clearinghouse (DEC), dating back to 1982. The gap between the red and blue lines indicates the extent to which documents coded as evaluations are not actually evaluations. As the figure indicates, the gap between documents coded as evaluations and documents verified as evaluations has narrowed in recent years, which indicates that DEC coding is becoming more accurate.

^{*}Figure 1 Data Sources:

1982–2012: USAID's Development Experience Clearinghouse.

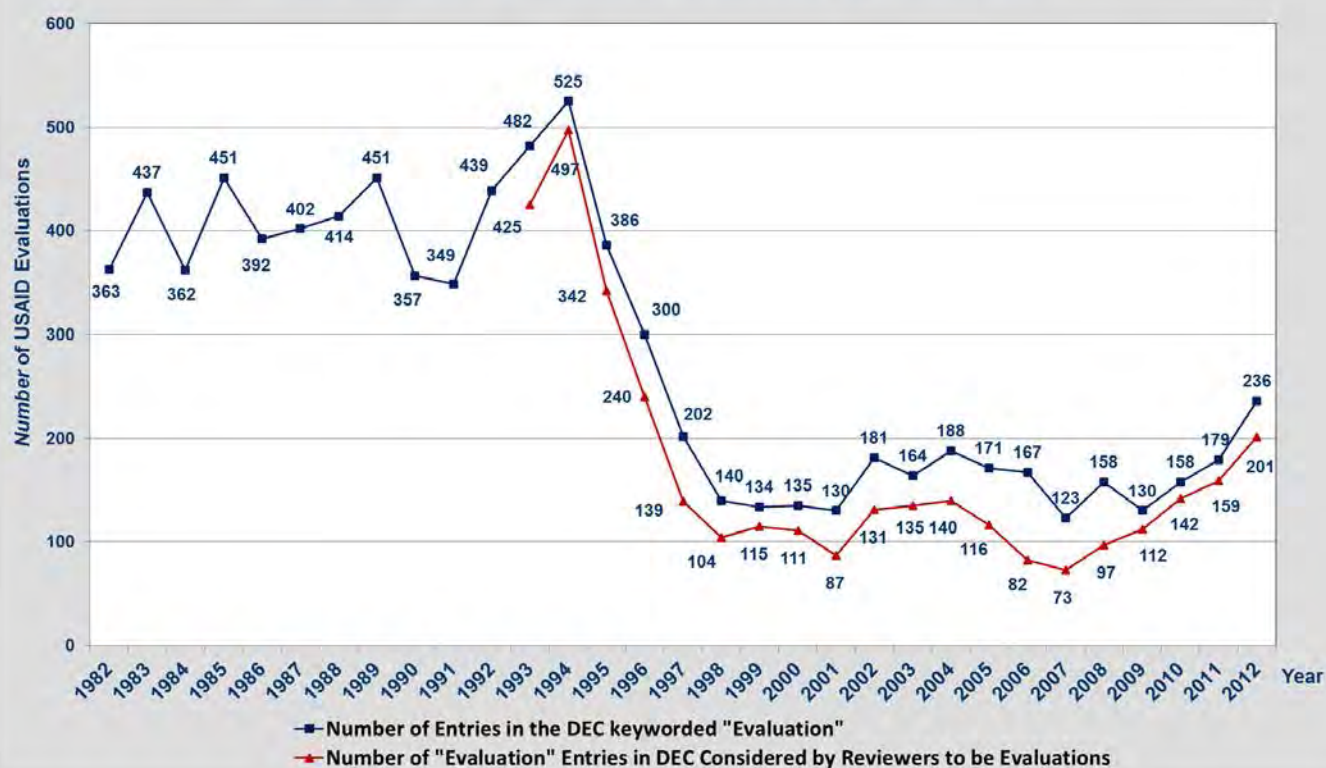
1993–97: Katrina Greene, "Narrative Summary of FY97 Evaluations," SAID/R&Rs, 7 January 1999.

1998: Cynthia Clapp-Wincek and Richard Blue, Evaluation of Recent USAID Evaluation Experience, USAID/PPC/CDIE Working Paper 320, 2001.

1999–2004: Janet Kerley USAID/CDIE.

2005–12: MSI.

[†]The decline pictured in Figure 3 paralleled USAID's 1995 elimination of a longstanding requirement for a midterm and final evaluation of every USAID-funded project. That same year saw the introduction of USAID's performance measurement system as an approach for obtaining progress information on programs and projects. In addition, after its creation in January 2006, the Office of the Director of Foreign Assistance in the Department of State, promoted an agenda that stressed monitoring. By 2008, interest in greater balance had re-emerged and both the F Bureau in State and USAID's Management Bureau had initiated small efforts to promote evaluation.

Figure 3. Trends in Number of Evaluations

Distribution by USAID Region or Bureau

In regional terms, the largest percentage of evaluations for the 2009–12 period came from Africa (38 percent), as shown in Figure 4. Africa's strong showing in the evaluation sample is consistent with its relatively large share of USAID funds. Within regions, certain countries produced a large share of their region's evaluations, including Afghanistan (83 percent) in the Afghanistan–Pakistan region (AfPak); Ethiopia (13 percent) and Uganda and South Sudan (9 percent each) in Africa (AFR); India (16 percent) and Indonesia (14 percent) in Asia; Kosovo (32 percent) and Georgia (17 percent) in the Europe & Eurasia (E&E) region; Nicaragua (17 percent), Columbia (14 percent), and Guatemala and Peru (12 percent each) in the Latin American and Caribbean (LAC) region; and Iraq (38 percent) and Yemen (15 percent) in the Middle East (ME).

Distribution by Sector or Topic of the Program or Project Evaluated

As shown in Figure 5, health program and project evaluations accounted for 29 percent of the evaluations examined in this study, reflective of USAID's significant investments in this sector. Health evaluations accounted for an even higher proportion (38 percent) in the 2005–08 meta-evaluations. Other shifts between the current meta-evaluation and previous studies include an increase in the percentage of Democracy and Governance (DG) evaluations from 13 percent to 23 percent, with a large share of these (26 percent) coming from the E&E region, and a small portion representing disaster recovery and food aid programs. A larger share of evaluations in the sample was also found for Economic Growth (EG) projects and programs, which rose from 11 percent to 20 percent, and includes a relatively large group of evaluations from Asia.

Figure 4. Geographic Distribution of 2009–12 Evaluations

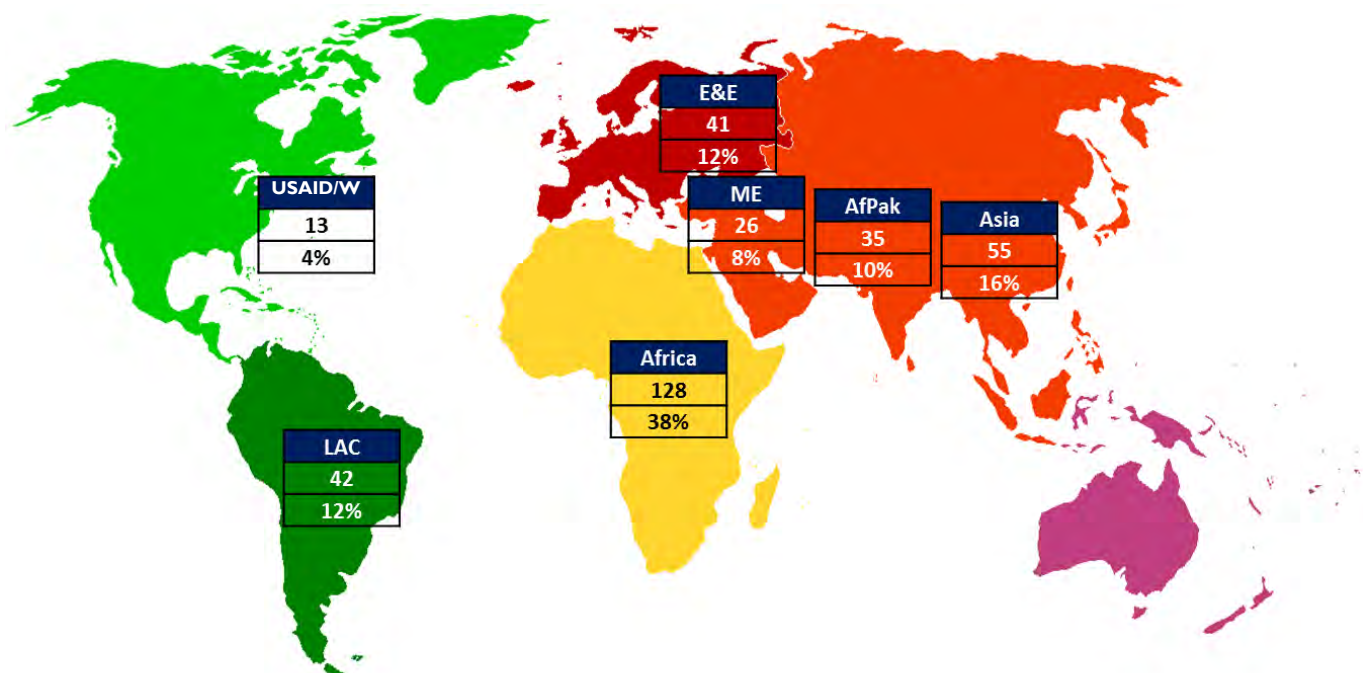
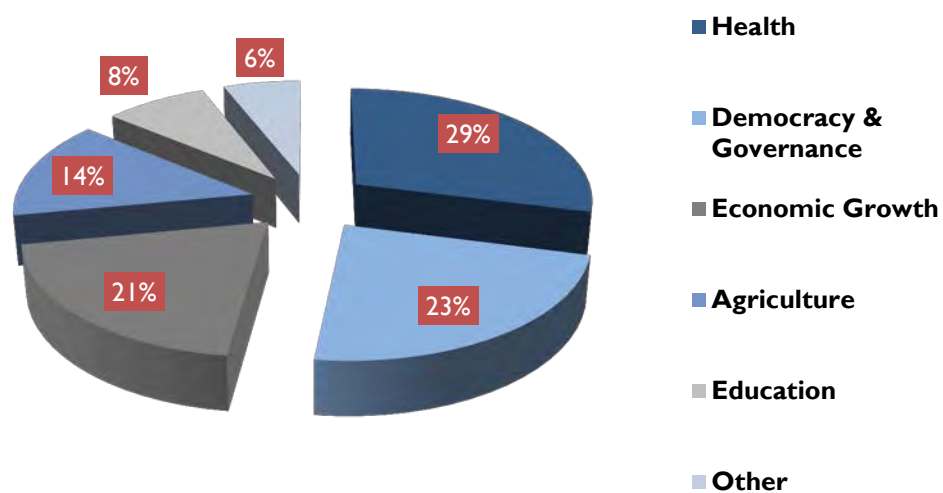


Figure 5. Sector Representation in the Meta-Evaluation Sample
(N = 340)



Scope or Scale of the Evaluation

Most USAID evaluations (76 percent) focused on a single project in a single country. This percentage parallels a finding from the previous (2005–08) meta-evaluation, and was fairly constant at this level across USAID regions and sectors. Among the 82 evaluations that did not concentrate on a single project, 82 percent were program-level evaluations in a single country. Evaluations with a larger scope of this sort can encompass all of the elements of a sector strategy focused on a specific Development Objective (DO). Other evaluations USAID conducted during this period included evaluations of global projects managed from USAID/Washington (USAID/W) and projects managed at the regional level.

Evaluation Timing

There was a roughly equal division between evaluations undertaken during project or program implementation (47 percent) and those undertaken toward or at the end of a program or project (45 percent). These figures are roughly equivalent to those for the 2005–08 meta-evaluation. That study categorized 46 percent of the evaluations it rated as being formative, which is roughly consistent with “during implementation,” and 43 percent as being summative, which is equivalent to taking place at or near the end of a program or project. In addition to evaluations undertaken during implementation or near the end of a project, 8 percent of the 2009–12 evaluations were undertaken on an ex-post basis, meaning after USAID funding terminated.

Type of Evaluation

In 2011, USAID’s Evaluation Policy updated its typology of evaluations. The current typology includes performance evaluations and impact evaluations. Performance evaluations are intended to address a wide range of questions of interest to USAID managers, including questions about project adherence to design-stage plans; overall performance relative to expectations; project efficiency or cost-effectiveness; the sustainability of services and benefits initiated under USAID projects; and whether projects and programs enhance gender equality. These are types of evaluation questions on which performance evaluations can produce high quality evidence, though not all do.

Impact evaluations, the second type of evaluation, concentrate on the effects of defined interventions inside of programs and projects. This is particularly true for innovative interventions that are the focal points of pilot projects and may be under consideration for scaling up. Impact evaluations involve formal comparisons between individuals, or other units that receive USAID assistance, and comparison groups that are constructed to demonstrate what would have occurred in the absence of USAID’s intervention.

Of the 340 evaluations examined in the meta-evaluation, 329 (97 percent) were performance evaluations and 11 (3 percent) were impact evaluations, which are the newer of these evaluation types and may take several years to complete. This balance appears to be consistent with USAID staff statements suggesting that roughly 90 percent of USAID evaluations are likely to be performance evaluations, while around 10 percent will employ more rigorous impact evaluation techniques.

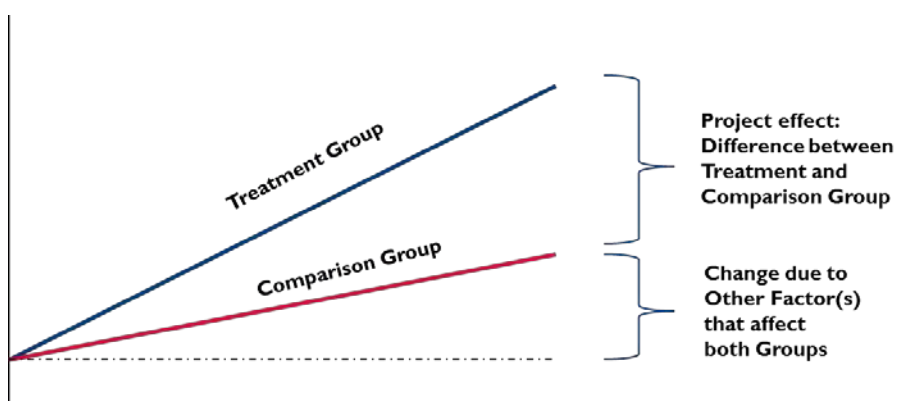
In addition to the 11 impact evaluations found in the study sample, the meta-evaluation found that 83 performance evaluations also addressed questions



On February 21, 2012, USAID released the final External Evaluation of the first five years of PMI.
Photo credit: Maggie Hallahan

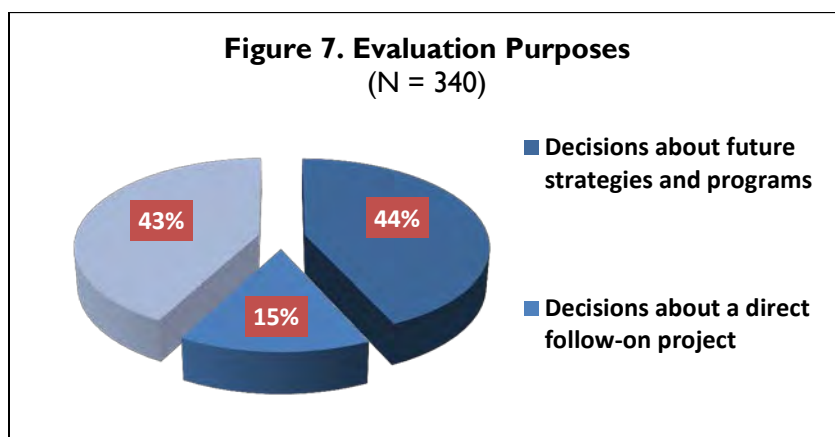
about cause-and-effect relationships. Some of these evaluations identified designs that are generally referred to as being non-experimental, while others simply described their data collection and analysis methods. A non-experimental design is appropriate when it is impossible to construct a comparison group, as is often the case for evaluations of policy interventions that affect whole populations. An interrupted time series is a common non-experimental design choice in such situations. Other non-experimental designs are conceptually akin to court processes that eliminate alternative possible causes and yield evidence consistent with the “beyond a reasonable doubt” standard. Performance evaluations that apply these approaches add value when USAID already knows, from an impact evaluation or other sources, that treatment group outcomes were significantly better than those for control groups, but also learned that both groups improved as a function of something else going on in their environment, as Figure 6 illustrates. Being able to determine, using non-experimental techniques, what other factors also contributed to improved outcomes can be as important for scaling up as knowing the effect size for an innovative intervention.

Figure 6. Non-Experimental Methods Can Help Identify Other Causes



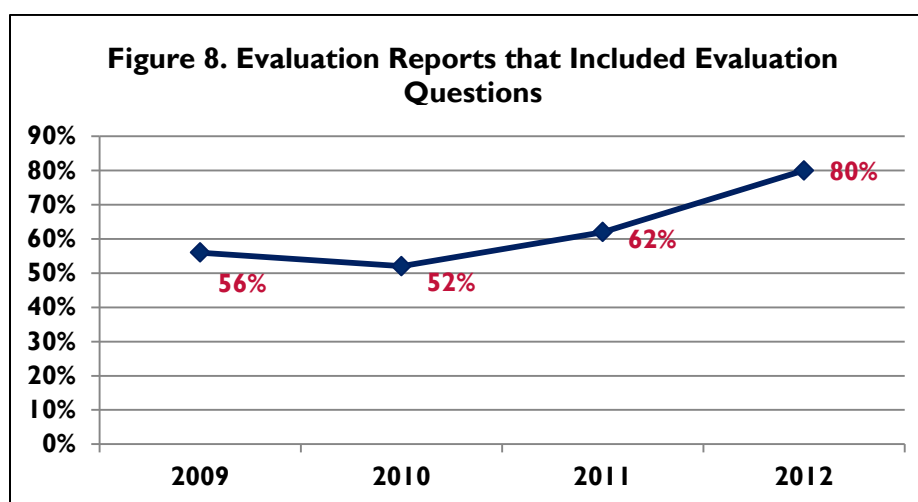
Evaluation Purpose

The most common management purposes for evaluations, shown in Figure 7 below, included providing evidence to inform decisions about future strategies or program designs (44 percent). The next largest cluster (43 percent) was focused on improving existing projects or programs. Only a small portion of evaluations were undertaken for the purpose of obtaining information that would be used to design direct follow-on projects. This mix of evaluation purposes appears to blend well with USAID program cycle requirements for considering the results of past evaluations when developing a Country Development Cooperation Strategy (CDCS) or designing a new project.



Presence of Evaluation Questions

Evaluations within and beyond USAID, have increasingly concentrated on specific questions raised by managers rather than on statements of issues or objectives. At USAID specifically, the percentage of evaluations that included or identified the existence of specific questions that teams were asked to address rose from around 56 percent in 2009 to 80 percent in 2012, as Figure 8 shows. At the start of the study period, it was much more common to find that evaluations focused on issues or objectives than at the end of the study. This is possibly a function of USAID’s guidance which states, in ADS 203.3.1.4, that evaluation SOWs should “identify a small number of key questions and specific issues answerable with empirical evidence.” In contrast, USAID’s Evaluation Policy and its evaluation training courses concentrate on questions as the starting point for an evaluation.



Number of Evaluation Questions

As previously noted, USAID evaluation guidance, since 2008 or earlier, has stressed that evaluations should address a “small number” of questions. MSI’s experience with evaluation courses funded by USAID has shown repeatedly that participants in those courses interpret the term “small number” as meaning 10 or fewer. Accordingly, the meta-evaluation counted the number of evaluations that met this standard. What it found was that over the four-year period 43 percent of evaluations included 10 questions or fewer. For 2012 alone, the percentage that met this standard rose to 52, while nearly half of all evaluations that included questions identified a larger number, as Table 2 shows.

Table 2. Number of Questions Asked in Evaluations
(N = 206)

Number of Evaluation Questions	Four-Year Average (2009–12)	2012 Percentages
1–10	43%	52%
11–20	29%	28%
21 or More	28%	20%

USAID PPL/LER recently released a How-To Note on preparing evaluation SOWs that suggests that evaluations should be asked to address three to five evaluation questions. Had that standard been used in the meta-evaluation, an even smaller percentage of the evaluations review would have been in compliance with USAID guidance.

On this issue, it is notable that attention to the number of evaluation questions included in SOWs emerged as an issue in earlier meta-evaluations. Those studies identified large numbers of questions in evaluation SOWs—often 25 or more—as one reason why USAID evaluation quality appeared to be relatively low. In this regard, it is noteworthy that evidence from this meta-evaluation, discussed under Question 3 below, was not able to demonstrate an association of this nature.

Evaluation Cost and Duration

Meta evaluations produced in the 1980s included information on evaluation cost and duration. They obtained this information from the face-sheet USAID Missions used when submitting their reports to the DEC (USAID FORM AID 1330–5) and shown later in the report. USAID appears to have ceased using this form around 1995 when other reengineering changes were introduced. Data on these variables are no longer readily available for USAID evaluations, thus this meta-evaluation and other recent meta-evaluations do not include detailed information on cost and duration. This precluded, among other things, an examination of the relationship between evaluation cost and evaluation scores under Question 3 below. With USAID’s assistance, MSI nevertheless attempted to obtain cost and duration data for evaluations completed between 2009 and 2012. Despite best efforts, obtaining reliable data on these variables proved impossible. MSI was, however, able to obtain qualitative data concerning evaluation cost and duration from its recent Team Leader Perceptions Survey and from group interviews.

Among 23 evaluation team leaders who responded to a question about evaluation resources, 40 percent indicated that the funds available for recent evaluations appeared to have declined when compared with earlier evaluations they had conducted. Comparatively, 36 percent said that resources were roughly the same and 16 percent reported that they felt resources had increased. Similarly, in small group interviews, one regional office representative and one firm indicated that they had seen little change in the size of evaluation budgets, while four other firms and a different regional office representative indicated that they had seen some larger budgets but that even those were still insufficient to cover the work involved, as that too had increased. Linking costs to study designs, one regional evaluation representative indicated that evaluation budgets do not reflect the number of questions asked. Another regional office representative added that, to some degree, the issue of evaluation budgets is being driven by the Office of Acquisition and Assistance (OAA), which always chooses the offer with the lowest cost. In turn, that regional office representative said, firms are beginning to respond to that formula and make offers that may be priced below the budget level actually needed to complete the work.

On the amount of time allotted for evaluations, data from the Team Leader Perceptions Survey, show that 36 percent thought less time was allocated for conducting the most recent evaluation than had been provided previously. While 40 percent indicated that they viewed duration as having remained roughly the same over time, the remaining 16 percent reported that more time was being allocated to conduct evaluations. Team leaders also indicated in their narrative responses they see a relationship between the time available for an evaluation and its quality.

B. Evaluation Quality Ratings

The overall picture of evaluation quality at USAID that emerges from this meta-evaluation is one of ongoing improvement over the four-year period of 2009–12. Not only has the number of evaluations completed per year increased, but the quality of USAID’s evaluation reports also has improved to the point that a statistically significant difference could be detected between evaluations completed in 2009 and those USAID produced in 2012. Notably, most of this improvement occurred between 2010 and 2012.

While this portrait is largely positive, the study also identified evaluation quality factors, or standards, that USAID evaluation reports do not yet meet. On several core evaluation quality standards—such as clear distinctions among evaluation findings, conclusions, and recommendations—performance was found to be below USAID standards. Other significant deficiencies included the small percentage of evaluations that indicated that an evaluation specialist was a member of the evaluation team, which USAID has required for the better part of a decade, and low ratings on the presence of sex-disaggregated data at all results levels—not simply for input level activities. In addition, low ratings were found for several evaluation standards introduced in the 2011 Evaluation Policy, but this may simply reflect slow uptake or a lack of awareness of these standards.

The study used a 37-point evaluation quality checklist to assess whether, and to what degree, evaluation reports met USAID's quality standards on factors determined primarily by evaluation teams, or in partnership with USAID (see Annex C for the checklist). Information from this checklist served as an important source of evidence for MSI's answers to three meta-evaluation questions discussed below.

Question 1. To what degree have quality aspects of USAID's evaluation reports, and underlying practices, changed over time?

Between 2009 and 2012 USAID evaluation reports realized net gains on 25 (68 percent) of 37 evaluation quality factors on the study rating checklist. Of these 25, 11 realized a net increase of 10 percentage points or more, making it unlikely that differences were simply small annual fluctuations that might not be indicative of real change. The 11 quality factors that registered net gains of 10 percentage points or more are listed in Table 3. While improvements on each of these factors was strong, Column 4 shows that, as of 2012, there were only four evaluation quality factors in this group on which 80 percent or more of the evaluations completed that year met USAID standards.

Table 3. Quality Factors With the Most Improvement Between 2009 and 2012

Evaluation Report Quality Factors		2009–12 Net Change	Percentage Rated Positively in 2012
#	Description		
Net Improvement of More Than 10 Percent on These Quality Factors Between 2009 and 2012			
6	Questions in report same as in SOW	57%	69%
33	SOW is included as a report annex	29%	74%
16	Study limitations were included	26%	64%
35	Annex included data collection instruments	25%	81%
12	External team leader	19%	82%
30	Recommendations—specific about what is to be done	19%	77%
18	Evaluation questions addressed in report (not annexes)	15%	74%
15	Report indicated conflict-of-interest forms were signed	12%	12%
22	Findings supported by data from range of methods	12%	80%
4	Management purpose described	11%	81%
23	Findings distinct from conclusions/recommendations	11%	48%

Not all evaluation quality factors improved over the study period. Quality ratings on 11 other quality factors (29 percent) declined between 2009 and 2012. In most instances these declines were modest (–1 percent to –3 percent), and not necessarily a matter for concern, particularly if their 2012 ratings in Column 4 were high. But for three factors the decline was steeper, as Table 4 shows.

Table 4. Factors With More Than a 1 Percent Decline in Quality

Evaluation Report Quality Factors		2009–12 Net Change	Percentage Rated Positively in 2012
#	Description		
Net Decline of More Than 1 Percent on Quality Factor Ratings Between 2009 and 2012			
24	Findings are precise (not simply “some, many, most”)	–7%	67%
19	Reason provided if some questions were not addressed	–11%	9%
11	Data analysis methods linked to questions	–13%	19%

Table 5 below provides year-by-year ratings on all 37 evaluation quality factors examined. They are presented in an order that follows USAID’s outline for an evaluation report. Also included is information on net change in the number of evaluations that addressed ten or fewer questions.

Table 5. Evaluation Quality Ratings for all Factors*

Evaluation Report Quality Factors (Full List)		Four-Year Average	Annual Ratings by Quality Factor				2009–12 Net Change [†]
#	Description		2009	2010	2011	2012	
1	Executive summary mirrors critical report elements	45%	41%	32%	63%	45%	5
2	Project characteristics described	90%	90%	87%	90%	91%	1
3	Project “Theory of Change” described	74%	77%	71%	75%	74%	–3
4	Management purpose described	80%	70%	81%	86%	81%	11
5	Questions were linked to purpose	99%	100%	97%	100%	98%	–2
6	Questions in report same as in SOW	50%	12%	50%	39%	69%	57
7	Written approval for changes in questions obtained	6%	8%	0%	4%	12%	4
8	Data collection methods described	90%	92%	80%	92%	96%	4
9	Data collection methods linked to questions	23%	17%	21%	29%	22%	5
10	Data analysis method described	33%	34%	25%	37%	34%	0
11	Data analysis methods linked to questions	31%	32%	40%	37%	19%	–13
12	External team leader	71%	64%	79%	57%	82%	18
13	Report said team included an evaluation specialist	13%	15%	12%	8%	19%	4

*This includes 37 factors on the meta-evaluation evaluation quality factor checklist plus an unnumbered factor that focused on whether the number of questions an evaluation addressed was 10 or fewer. Notably, two quality factors were dropped from the analysis when interrater reliability assessments indicated that differences between raters on these factors made them unreliable.

[†]In percentage points.

META-EVALUATION OF QUALITY AND COVERAGE OF USAID EVALUATIONS 2009–12

Evaluation Report Quality Factors (Full List)		Four- Year Average	Annual Ratings by Quality Factor				2009–12 Net Change [†]
#	Description		2009	2010	2011	2012	
14	Evaluation team included local members	30%	33%	25%	26%	35%	2
15	Report indicated conflict-of-interest forms were signed	3%	0%	0%	1%	12%	12
16	Study limitations were included	51%	38%	34%	62%	64%	26
17	Report structured to respond to questions (not issues)	45%	47%	36%	47%	51%	4
18	Evaluation questions addressed in report (not annexes)	62%	59%	71%	44%	74%	15
19	Reason provided if some questions were not addressed	9%	21%	8%	3%	9%	–12
20	Social science methods (explicitly) were used	77%	81%	64%	78%	84%	3
22	Findings supported by data from range of methods	74%	68%	71%	74%	80%	12
23	Findings distinct from conclusions/recommendations	41%	37%	42%	37%	48%	11
24	Findings are precise (not simply “some, many, most”)	66%	74%	64%	63%	67%	–7
25	Unplanned/unanticipated results were addressed	15%	15%	11%	19%	14%	–1
26	Alternative possible causes were addressed	10%	10%	8%	11%	10%	0
27	Evaluation findings sex disaggregated at all levels	20%	23%	15%	22%	22%	–1
28	Report discusses differential access/benefit for men/women	32%	42%	27%	23%	40%	–2
29	Recommendations—not full of findings, repetition	59%	58%	56%	56%	64%	6
30	Recommendations—specific about what is to be done	72%	58%	79%	72%	77%	19
31	Recommendations—specify who should take action	49%	43%	45%	63%	45%	2
32	Recommendations—clearly supported by findings	80%	80%	76%	83%	79%	–1
33	SOW is included as a report Annex	58%	45%	38%	68%	74%	29
34	Annex included list of sources	78%	84%	68%	80%	83%	–1
35	Annex included data collection instruments	61%	56%	49%	55%	81%	25
37	Statements of differences included as an Annex	4%	3%	2%	4%	7%	4
38	Report explains how data will transfer to USAID	2%	0%	1%	2%	5%	5

Evaluation Report Quality Factors (Full List)		Four- Year Average	Annual Ratings by Quality Factor				2009–12 Net Change [†]
#	Description		2009	2010	2011	2012	
39	Evaluation SOW includes Evaluation Policy Appendix I	6%	0%	0%	9%	8%	8
N/A	Number of evaluation questions was 10 or fewer	32%	49%	36%	20%	29%	–20

Only two of the quality factors that improved between 2009 and 2012 did so in a stepwise, or linear, fashion (Factors 15 and 38).^{*} Most factors that improved did so by rising and falling between the start and end of the study period, ending higher in 2012 than in 2009. Also of note, there were four quality factors with a noticeable upward shift in the final two years of the study period, and all of these instances involved new requirements introduced in 2011 (Factors 9, 22, 33, and 39).

Factor Quality Ratings Relative to Standards

Short paragraphs below examine key findings for specific evaluation quality factors. Part 2 of this volume provides a more detailed review of each factor, and readers specifically interested in quality ratings on a regional or sector basis, or for USAID Forward evaluations are encouraged to examine that section of the report.[†]

Executive Summary

The number of USAID evaluations that include an executive summary doubled in the 30 years since USAID began conducting meta-evaluations. This is a positive finding because busy USAID managers who are interested in what an evaluation found may have only enough time to read the executive summary. However, only roughly half of all executive summaries conformed to USAID's expectation that they would faithfully mirror all key aspects of the evaluations they summarized. Many of the executive summaries analyzed had left out one or more key evaluation report elements. Others included new information not previously discussed in a report.

Introductory Information

USAID evaluations received high ratings on the inclusion of project or program background information (91 percent of 2012 reports were scored positively); the presentation of a management purpose (81 percent of 2012 evaluations met USAID expectations); and having a clear relationship between the evaluation's management purpose and questions (98 percent of 2012 evaluations scored positively). Somewhat less positive was a quality rating factor on the inclusion of the project's or program's theory of change, or what USAID calls Development Hypotheses, where only 74 percent of 2012 evaluations earned a positive rating. Inclusion of the program's or project's theory of change is important because it explains to readers of an evaluation report how USAID expected its efforts would bring about change and what results it sought. Without a theory of change it may be difficult for a reader to interpret findings, conclusions, and recommendations presented in an evaluation.

Team Composition

Findings on team composition indicate a low level of compliance with USAID's requirement that every evaluation team include an evaluation specialist.

^{*}The numbering sequence on the evaluation quality factor checklist runs to 39 items. The actual total is 37, as 2 items were removed when interrater reliability checks identified problems with rating consistency.

[†] Ratings on a regional basis can be found in Table 63 on page 98). Average ratings by sector are shown in Table 64 on page 100), and USAID Forward ratings on factors are provided Table 66 on page 102.

USAID evaluation guidance includes three imperatives with respect to team composition. It requires that 1) USAID evaluation team leaders be external—having no relationship to USAID or the program or project being evaluated—and that 2) every evaluation team include at least one evaluation specialist. USAID’s Evaluation Policy also 3) encourages the inclusion of local professionals on USAID evaluation teams, including as team leaders. The study reviewed USAID evaluations to determine the extent to which they reported on team members with these characteristics.

Table 6 shows the extent to which evaluation reports described these three types of team members. The percentage of external team leaders rose from 55 percent in 1983 to 61 percent in 1989–90, and to 71 percent for the four-year study period. For 2012, the meta-evaluation found that 83 percent of team leaders were external to the Agency. This figure represents an increase over the four-year average of more than 10 percentage points.

Table 6. Team Members Identified in Evaluation Reports

Team Composition	Four-Year Average Percentage (2009–12)	2012 Percentages
External team leader	71	83%
Evaluation specialist	14	19%
Local team member(s)	29	35%

At the same time, the study showed that only 19 percent of 2012 evaluation reports said that teams included an evaluation specialist. This figure is much lower than the percentage of evaluations reporting external team leaders. The small percentage of evaluations that indicated the presence of an evaluation specialist is somewhat surprising, given that since 2008 or before, USAID ADS 203 has required that evaluation teams include an evaluation specialist. As discussed further under Question 3 below, the presence of an evaluation specialist on a team is positively associated with strong overall evaluation quality.

USAID evaluation guidance also encourages Missions to include country partner representatives on evaluation teams. As Table 6 shows, 35 percent of evaluations completed in 2012 reported that they did this. The status of this factor is important because the involvement of local professionals is considered an effective way to build evaluation capacity in partner countries—a goal to which USAID subscribes.

Evaluation Methodology

While reviews of existing performance data and interviews are the methods most frequently used to obtain data to answer evaluation questions, a comparison with previous meta-evaluations showed that the use of more formal social science methods such as surveys, focus groups and structured observation is higher in 2009–12 than was the case in earlier years. Surveys, for example, were used to collect data in 10% of evaluations in 1997–98 compared to 35% of 2009–12 evaluations.

Team leaders of recent USAID evaluation who responded to a survey report that in their experience, evaluation methods are often based (60 percent of the time) on suggestions in evaluation SOWs that also encourage an evaluation team to offer their own comments and ideas. The result is a methodology to which both USAID and an evaluation team contribute.

Ninety percent of USAID evaluation reports describe the methods they used to collect data. A lower proportion (33 percent) describe the methods they use to analyze data. While USAID’s 2012 How-To Note on evaluation reports recommends that data collection methods be described on a question-by-question basis to ensure that the best possible methods are used in each case, few evaluations in the

study sample did this. Presentations of evaluation methods on a question-by-question basis were found in 20 percent of recent evaluations.

Among the data collection sources and methods used, existing performance information and document reviews were the most frequently used resources. Primary data were collected mainly from key informant interviews, individual interviews, and unstructured observation. Less frequently cited were more formal social science research methods such as surveys, focus groups, and structured observation techniques, as Table 7 shows.

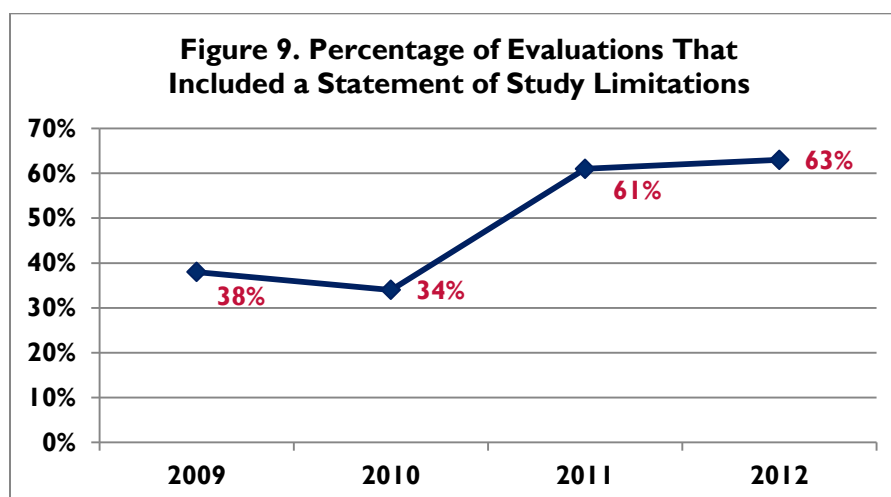
Table 7. Data Collection Methods Used in 2009–12 Evaluations

Data Collection Methods Used	Percentage of Evaluations That Demonstrated Use of the Method
USAID performance data	84%
Document review	81%
Key-informant interviews	72%
Individual interviews	54%
Unstructured observation	45%
Survey	35%
Focus group	29%
Structured observation	8%
Group interview	9%
Instruments (e.g., scale)	3%
Community interview	1%

On the data analysis side, descriptive statistics were identified as being used in 63 percent of 2009–12 evaluations. Content analyses of qualitative data were reported in 27 percent of these evaluations, and inferential statistics, such as a test between the means for two groups, were cited in 9 percent of evaluations.

Study Limitations

Performance improved dramatically in the last two years of the study period on the inclusion of a statement of study limitations in evaluation reports. This is one of four evaluation quality factors where a spike in performance was noted for 2011 and 2012. As Figure 9 shows, ratings on this factor rose from around 35 percent in 2009-10 to around 60 percent in 2011-12. Including study limitations in an evaluation report improves the report's professionalism by ensuring that readers are aware of any reservations they should have concerning the quality of the data on which an evaluation's conclusions and recommendations are based.



Structure for Reporting Evaluation Findings

Questions were used as the organizing structure for reporting evaluation findings in 54 percent of evaluations that were asked to address specific questions. A question-and-answer structure of this sort helps managers quickly identify how an evaluation's results relate to decisions a Mission is facing. Of those that did this, the percent that addressed exactly the same list of questions as was found in accompanying evaluation SOWS increased over the study period, rising from 12 percent of reports in 2009 to 69 percent in 2012. Answering each evaluation SOW question in this manner is precisely what USAID's Evaluation Policy demands.

Presentation of Evaluation Findings

USAID's evaluation policy expects that evaluation findings will be presented in a professional manner. The meta-evaluation used four quality factors to assess professionalism in this area and found on average:

- 77 percent of the evaluations used social science methods to arrive at findings. This finding aligns with standards included in USAID's Evaluation Policy.
- 74 percent derived study findings from the full range of methods the team demonstrated that it used. This finding suggests that evaluation teams fully analyzed the information they collected.
- 66 percent scored positively on whether findings, particularly quantitative findings, were presented precisely, meaning numerically rather than as broad statements such as "some" respondents said, or "most" farmers participated.
- 41 percent of evaluations adequately distinguished among findings, conclusions, and recommendations. On an annual basis, ratings on this factor rose from 37 percent in 2009 to 48 percent in 2012.

Among these four ratings, the blurring of findings, conclusions, and recommendations is particularly problematic for readers of evaluation reports. In 2012, fewer than half of the evaluations reviewed met USAID standards for appropriately separating these evaluation elements by introducing shifts from findings to conclusions, or conclusions to recommendations, in the text of a report. Reports that rely on sweeping statements about what was found ("most farmers said") are also problematic, as the evaluation audience does not know whether "most" means 51 percent or 99 percent; precision on such points can affect the priority USAID assigns to taking action on evaluation findings.

Gender Considerations in Evaluations

USAID's March 2012 Gender Equality and Female Empowerment Policy commits USAID to learning, through evaluations, about the effects of agency programs from a gender perspective and to designing programs and projects with their gender-specific impacts clearly in mind. The inclusion of two gender-specific evaluation quality rating factors in the meta-evaluation not only checked on the responsiveness of evaluations to past USAID guidance, but also established a baseline for assessing future improvements in response to USAID's updated gender policy.

Evaluation quality ratings from the meta-evaluations were low on both of the factors assessed. MSI found that only 20 percent of the evaluations included sex-disaggregated data at all results levels where such data could, in principle, have been collected. Further, only 32 percent of evaluations over the four-year period included at least some mention of gender differential aspects of a project. MSI undertook a post-rating review of how gender issues were handled in evaluation reports. This review indicated that in most instances discussions of gender differential effects were based on limited data, including anecdotes, rather than on systematic data collection and analysis.

USAID Evaluation Policy Aims to Professionalize Evaluation Practices

Evaluation "Good Practices" emphasize USAID's reliance on facts in reaching conclusions and recommendations:

- Use of data collection and analytic methods that 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.
- Application and use to the maximum extent possible of social science methods and tools that reduce the need for evaluator-specific judgments.

Inclusion of Findings on Broad Evaluation Concerns

A distinguishing feature of evaluations is, in principle, the holistic way in which evaluation teams examine the results of projects and programs. Thus, for example, one might expect a performance monitoring report to discuss only the planned results of a project, or look only at the project intervention with one or two critical assumptions when discussing why outputs and outcomes are or are not being produced. An evaluation differs conceptually from performance monitoring in that the former has an implicit mandate to examine the "worth" of a project and its full range of results. Two evaluation quality factors in the meta-evaluation served as proxies for testing whether USAID evaluations are holistic in this sense. The first of these factors focused on whether evaluations reported on unplanned effects, which 15 percent of the 340 evaluations in the study sample did. The second factor concentrated on whether evaluations discussed alternative possible causes of observed results, or causes in addition to USAID interventions that might be contributing to results. Only 10 percent of the evaluations were found to have done this.

Presentation of Evaluation Recommendations in Reports

Recommendations play a critical role in evaluations. The World Bank, for example, reported in its 2008 annual report that an internal study had discovered that poorly framed recommendations were less likely to be implemented than recommendations that were stated in clear and actionable terms. In this meta-evaluation, four quality factors were used to rate evaluation recommendations. USAID evaluations performed reasonably well on one of these factors but on others they rated below USAID's expectations.

- 72 percent of the evaluations were rated as having recommendations that are specific about what is to be done.
- 60 percent of the evaluation recommendation sections were found to be clearly supported by findings.

- 59 percent were rated positively on being distinct from other aspects of reports and not unnecessarily repetitive of earlier report sections.
- 49 percent were rated positively for having clearly indicated who or which organizations need to take action on the evaluation's recommendations.

Among these ratings, the fact that only 60 percent of evaluation recommendations were clearly supported by findings is particularly problematic, as it suggests that in 40 percent of USAID evaluations at least some recommendations are simply appearing without demonstrating a foundation in known facts about the program or project being evaluated. One of the aims of USAID's list of good practices on page 9 of its Evaluation Policy is to limit this type of practice.

Annexes

Evaluation report annexes include numerous elements that contribute to distinguishing highly professional studies from those carried out in a more ad hoc manner. Accordingly, several of the evaluation quality checklist items focused on this area. A total of six annex-related factors are discussed briefly below. Three of these annexes are associated with longstanding requirements.

- **Evaluation SOW.** This required annex both guides an evaluation team and serves as the framework for reviewing evaluations to determine whether USAID received what it contracted for, or otherwise requested. Ratings for this factor improved over the study period, rising from 45 percent in 2009 to 74 percent in 2012. Data from earlier meta-evaluations show that 74 percent compliance in 2012 is not actually a gain but rather a return to levels recorded in both 1983 and 1989–91.
- **List of Sources.** Seventy-nine percent of evaluations over the four-year study period included a list of sources.
- **Evaluation Instruments.** Sixty-one percent of evaluations provided this expected annex, including multiple instruments in some cases.
- **Conflict of Interest Statements.** This is a new requirement, and 4 percent of evaluations in the study sample were found to have included this annex.
- **Data Transfer to USAID.** This is another new requirement, and 2 percent of evaluations complied.
- **Statement of Differences.** This is an annex that is used only when needed. It is not expected to improve in the same manner as other evaluation quality factors. Only 4 percent of the study sample of 2009–12 evaluations included this annex.



Photo credit: USAID/Haiti

Question 2. At this point in time, on which evaluation quality aspects or factors do USAID’s evaluation reports excel and where are they falling short?

On half of the evaluation quality factors rated, USAID evaluations fell into “good” and “fair” clusters indicating that USAID standards on these factors are understood and efforts to achieve them are being made. USAID evaluation reports were judged to have excelled on any evaluation factor and were coded “good” where 80 percent of its 2009–12 evaluations met a given USAID’s quality standard, as Table 8 illustrates.

Table 8. Performance on Evaluation Factor Quality Ratings by Cluster

Percentage of Evaluations That Met USAID’s Quality Criteria in 2012		Evaluation Factors	
Cluster	Basis for Cluster	Number	Percentage
Good	80% of or more met quality criteria	9	24%
Fair	50% to 79% met criteria	11	29%
Marginal	25% to 49% met criteria	6	16%
Weak	Fewer than 25% met criteria	12	32%

In all, 24 percent of the 37 evaluation quality factors for evaluations completed in 2012 were coded “good.” As Table 9 shows, another 11 factors were rated as “fair,” meaning between 50 percent and 79 percent of USAID evaluations met USAID’s quality standard. On the remaining 18 factors, fewer than 50 percent of evaluations met USAID’s quality standard. These factors fell into clusters coded as “marginal” and “weak.” In six instances, factors rated in these clusters are new quality standards introduced in the 2011 Evaluation Policy, where uptake may not yet be complete. Three other factors—two in the “weak” cluster and one in the “marginal” cluster—are factors that do not necessarily apply to all evaluations, such as including a statement of differences. Weak or marginal scores on these indicators are not overly concerning for these reasons.

Table 9. Quality Factor Ratings for 2012

Evaluation Report Quality Factors (Full List)		Rated Positively in 2012	Factors Status in 2012
#	Description		
5	Questions were linked to purpose	98%	Good
8	Data collection methods described	96%	Good
2	Project characteristics described	91%	Good
20	Social science methods (explicitly) were used	84%	Good
34	Annex included list of sources	83%	Good
12	External team leader	82%	Good
4	Management purpose described	81%	Good
35	Annex included data collection instruments	81%	Good
22	Findings supported by data from range of methods	80%	Good
32	Recommendations—clearly supported by findings	79%	Fair
30	Recommendations—specific about what is to be done	77%	Fair
3	Project “Theory of Change” described	74%	Fair

Evaluation Report Quality Factors (Full List)		Rated Positively in 2012	Factors Status in 2012
#	Description		
18	Evaluation questions addressed in report (not annexes)	74%	Fair
33	SOW is included as a report annex	74%	Fair
6	Questions in report same as in SOW	69%	Fair
24	Findings are precise (not simply “some, many, most”)	67%	Fair
16	Study limitations were included	64%	Fair
29	Recommendations—not full of findings, repetition	64%	Fair
17	Report structured to respond to questions (not issues)	51%	Fair
23	Findings distinct from conclusions/recommendations	48%	Marginal
1	Executive summary mirrors critical report elements	45%	Marginal
31	Recommendations—specify who should take action	45%	Marginal
28	Report discusses differential access/benefit for men/women	40%	Marginal
14	Evaluation team included local members	35%	Marginal—May Not Apply
10	Data analysis method described	34%	Marginal
N/A	Number of evaluation questions was 10 or fewer	29%	Marginal
9	Data collection methods linked to questions	22%	Weak—New
27	Evaluation findings sex disaggregated at all levels	22%	Weak
11	Data analysis methods linked to questions	19%	Weak—New
13	Report said team included an evaluation specialist	19%	Weak
25	Unplanned/unanticipated results were addressed	14%	Weak—May Not Apply
7	Written approval for changes in questions obtained	12%	Weak—New
15	Report indicated conflict-of-interest forms were signed	12%	Weak—New
26	Alternative possible causes were addressed	10%	Weak—May Not Apply
19	Reason provided if some questions were not addressed	9%	Weak—Small N
39	Evaluation SOW includes Evaluation Policy Appendix I	8%	Weak—New
37	Statements of differences included as an annex	7%	Weak—Small N
38	Report explains how data will transfer to USAID	5%	Weak—New

Factors rated “marginal” or “weak” that were not new evaluation quality factors warrant USAID’s attention. Thus, in the paragraphs below, MSI reviews evaluation quality factors at the low end of these ratings. Performance on many of these factors can be improved by USAID staff for evaluations over which they have direct control.

Quality Factors in the Weak Cluster

Two “weak” quality factors cannot be easily explained as a function of the recency of the standards they represent or because they do not apply in all cases:

- **Evaluation Specialist Identified.** The low percentage of evaluation reports that indicated an evaluation specialist was a member of the team is problematic given that USAID has required the presence of such specialists since 2008 or earlier. In small group discussions with USAID staff, and conversations with PPL/LER, the lack of consensus about what the term *evaluation specialist* means may be an issue.
- **Sex-Disaggregated Data at All Results Levels.** While evaluations did relatively well on including sex-disaggregated data at the input–output level, on indicators such as a count of men/women

attending a training course, they often failed to meet this standard by neglecting to gather sex-disaggregated data on outcomes, such as the percentage of women/men who adopted a new technology (that was the object of their training program).

Quality Factors in the Marginal Cluster

Evaluation quality factors that fell into the “marginal” cluster are quality issues on which some guidance existed before the start of the meta-evaluation study period, but which were, in most instances, reinforced in USAID’s Evaluation Policy and associated guidance.

- **Executive Summary.** Fifty-nine percent of evaluations were rated as not presenting a complete and accurate summary of their evaluation. Missing elements were the most common problem.
- **Evaluation Questions.** Seventy-one percent of the evaluations that included evaluation questions identified more than 10 questions for the evaluation team to address, which was treated by the meta-evaluation as the upper limit with respect to USAID guidance on asking a “small number” of questions.
- **Team Composition.** While it is not necessarily expected that all USAID evaluations will include local team members, 65 percent of the evaluations failed to identify, and thus presumably lacked, local professionals on their evaluation teams.
- **Evaluation Methods.** Sixty-six percent of the evaluations failed to describe how evaluation data would be analyzed in their methods section, or in an annex on that topic.
- **Evaluation Findings.** Fifty-two percent of evaluations mingled findings, conclusions, and recommendations to some degree, or simply failed to make clear transitions when moving between these three elements of evaluation reports.
- **Gender Considerations.** Even though the bar was set low for the inclusion of information on gender-specific and gender-differential participation in, or results of, program and projects, 60 percent of the evaluations did not address these issues.
- **Recommendations in Evaluation Reports.** Fifty-five percent of the evaluations failed to specify who was expected to take actions on evaluation recommendations.



USAID Staff Efforts to Improve Evaluation Quality

Over the course of the meta-evaluation, MSI learned from USAID staff, evaluation team leaders, and organizations that undertake evaluations for USAID about aspects of the Agency’s evaluation practice that are improving, though not necessarily in all offices or Missions. These quality improvement and quality control efforts are worth noting in connection with evaluation quality factors in the “weak” and “marginal” clusters discussed. Improvements noted in small group meetings and the team leader survey are summarized below with respect to evaluation SOWS and evaluation reports.

Evaluation Statements of Work (SOWs)

In small group interviews with regional and technical bureau staff, as well as with firms carrying out evaluations for USAID, participants identified evaluation SOWs as a factor affecting evaluation quality. Six of the firms that participated in these interviews said that the overall quality of USAID evaluation SOWs has improved—they are clearer and generally better, with one technical office representative offering a similar perception. Three other firms, however, disagreed, saying that the evaluation SOWs they have seen recently are either roughly the same as in the past or possibly worse.

One technical representative pointed out that there is a difference between SOW quality in Washington and in the field, and that the improvements are taking place much more in Washington than overseas. Two other small group participants said that USAID's evaluation training courses are having a positive influence on the quality of evaluation SOWs. However, one firm commented that these trainings are insufficient and that SOW writers need to experience evaluations in the field to be able to write good ones. On this same point, five of the firms noted that the language being used in SOWs is often cut and pasted from the Evaluation Policy, or include “buzz words” that writers have heard but do not understand. This is demonstrated by the use of incorrect meanings for some terms or an apparent lack of understanding of the methodological implications of some of the Evaluation Policy's guidance. This, they said, leads to poor-quality SOWs and creates confusion in the firms about how to respond to such solicitations, both technically and financially.

Evaluation Management Practices

Participants in small group discussions also commented on USAID's processes for reviewing SOWs before releasing them. Regarding the SOW peer-review process taking place following USAID's release of the 2011 Evaluation Policy, a regional representative stated that the review process is generally accepted and appreciated, though there was much resistance at first. One regional bureau representative also stated that the SOW review process has increased SOW quality, although another regional representative said that, in the region's view, SOW quality may have decreased as a result of reviews. Four regional and two technical office staff representatives stated that the current process for reviewing SOWs does not provide enough direction on how, when, or by whom they should be done, leading to inconsistency in quality of reviews and reviews being done by people without enough evaluation knowledge.

Commenting on USAID evaluation planning and management processes more broadly, 46 percent of respondents to the Team Leader Perceptions Survey indicated that they are now being asked to prepare reports before initiating fieldwork for an evaluation. These reports, which the World Bank and other UN organizations call inception reports, involve one or two elements. The first element, which some team leaders reported having been asked to prepare, involved summarizing what is already known about evaluation questions from performance data and identifying gaps that remain. In addition, 71 percent of respondents indicated that in recent evaluations they have been asked to produce detailed evaluation designs, including sampling plans and instruments, before starting their fieldwork, which are also an element of evaluation inception reports required by other development agencies.

Additional evaluation management quality checkpoints and post-evaluation fieldwork were also identified as being required by USAID in recent evaluations. Responses from 23 of the 25 respondents (92 percent) to the Team Leader Perceptions Survey reported that they had been asked to provide a briefing on findings, conclusions, and recommendations before writing their draft reports. One firm participating in the small group interviews even commented on having seen requests for reports on “findings to date” during the evaluation field data period.

In addition, in the meta-evaluation's Team Leader Perceptions Survey, one question asked about the thoroughness with which recent evaluation reports have been reviewed, compared with earlier evaluation reports with which team leaders have been involved. Team leader responses indicate that in most cases the quality of reviews remains unchanged, but that some team leaders consider reviews to have become more rigorous, particularly with respect to the strength of the evidence needed to support evaluation conclusions and recommendations. Comments on evaluation report reviews were also offered in small-group meetings with USAID regional bureau staff, technical bureau staff, and firms that conduct evaluations. With respect to enhanced reviews of evaluation reports, one regional representative stated that the review of reports is done inconsistently by Missions, with some sending them to Washington and others not. In another session, three evaluation firms stated that it is clear that USAID has been using checklists when reviewing evaluation reports.

Commenting more broadly on these kinds of quality-control processes, three technical and regional representatives commented that the evaluation process at USAID is improving, while one regional bureau and two technical representatives said that this process improvement has not yet translated into an improvement in the quality of evaluation reports. Another regional bureau representative commented that there are now much timelier submissions of evaluation reports to the DEC, though this does not directly reflect evaluation quality itself.

Question 3. What can be determined about the overall quality of USAID evaluation reports and where do the greatest opportunities for improvement lie?

Through this question, USAID hoped to learn about interactions between and among quality factors and whether improvements in some factors might have multiplier effects, such that several quality factors would improve simultaneously. To address this question the meta-evaluation team created an overall evaluation score from 11 of the 37 factors included in the study's evaluation quality factor checklist. Two of the factors used were combined, allowing individual evaluations to receive ratings ranging from 0 to 10, as discussed further in Part 2 and Annex B.

Average Overall Evaluation Scores

The average overall evaluation score for the 340 USAID evaluations completed between 2009 and 2012 was 5.93 out of 10, while the mode was slightly higher at seven points. Average scores offer an easy way to compare groups of evaluations and determine whether differences between groups are statistically significant. Using t-tests to compare groups, the two most striking findings from these comparisons show the following:

- Average overall quality scores for 2012 (6.69) were significantly higher than the average scores for 2009 (5.56). This finding suggests that some set of intervening factors may explain why average scores are different in these two years, such as activities under USAID's evaluation quality improvement initiative, including new policy, trainings, and evaluation awards. The difference between years is not due to chance alone.
- Average overall scores for evaluations where an evaluation specialist was reported to be a team member (6.87) were higher than for those evaluations where the participation of an evaluation specialist was not reported (5.78). This finding was statistically significant at the .05, .01 and .001 levels. It suggests that evaluation specialists play a role that affects overall evaluation scores.

By contrast, differences between USAID Forward and non-USAID Forward evaluations completed between July 2011 and December 2012, or for evaluations with or without an external team leader, were not significant.

Associations Between Overall Scores and Individual Evaluation Quality Factor Scores That Are Not Components of the Overall Score

In addition to comparisons between averages for groups, chi-square tests that reveal the degree of association between factors were used to examine the relationship between overall scores and evaluation ratings on a variety of factors that were not used to create those overall scores.

Table 10 displays evaluation quality factors and other evaluation characteristics found to be statistically associated with overall scores, rank ordered by chi-square value, and showing associated levels of significance to the right. Both year and presence of an evaluation specialist are included in the analysis.

Table 10. Degree of Association Between Evaluation Scores and Other Evaluation Characteristics

#	Quality Rating Topic	Chi-Square Value	Significance Levels*		
			.05	.01	.001
	Year in which evaluation was completed	29.601	●	●	●
25	Unplanned/unanticipated results were addressed	20.682	●	●	●
9	Data collection methods linked to questions	19.567	●	●	●
15	Report indicated conflict-of-interest forms were signed	18.196	●	●	●
22	Findings supported by data from range of methods	17.364	●	●	●
	Sector associated with program/project evaluated	17.330	●		
13	Report said team included an evaluation specialist	14.674	●	●	●
	Purpose was to support design of future strategies, programs, projects	14.542	●	●	●
26	Alternative possible causes were addressed	13.720	●	●	
18	Evaluation questions addressed in report (not annexes)	13.346	●	●	
38	Report explains how data will transfer to USAID	11.984	●	●	
6	Questions in report same as in SOW	10.305	●		
17	Report structured to respond to questions (not issues)	10.193	●		
29	Recommendations—not full of findings, repetition	8.606	●		
14	Evaluation team included local members	7.344	●		

As Table 10 shows, the strongest association found between overall scores and other evaluation characteristics involved the year in which evaluations were completed. The fact that this factor has the strongest association with overall scores again suggests that evaluation-related changes at USAID, occurring over the 2009–12 time period, played an important role in improving evaluation quality.

Most of the 15 factors and characteristics that were found to be statistically associated with high overall scores are not “drivers” of evaluation quality but rather characteristics that reflect decisions and actions taken by USAID staff or evaluation teams. The level of attention given to these factors may reflect policy priorities. In contrast, other factors in this group may be “drivers” of quality ratings. The presence of



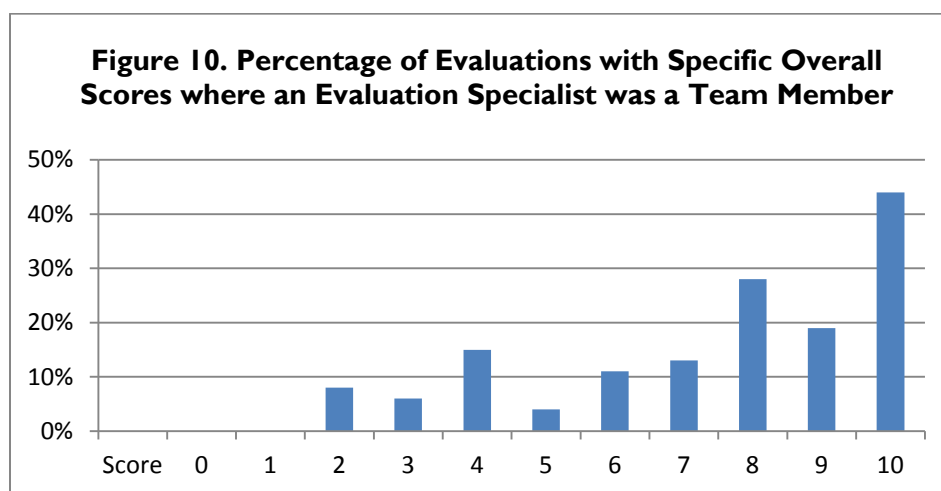
*For all numbered factors, the degrees of freedom in the chi-square test was 2, but for some of the unnumbered items included in this table the degrees-of-freedom number was higher. This explains why chi-square values that were roughly the same were not necessarily significant at all the same significance levels.

an evaluation specialist is potentially this type of factor. Similarly, the year in which the evaluation was completed—which is discussed above as a possible proxy for policy change—and sector, where differences in the level of experience with rigorous evaluation approaches may play a role may be “drivers” of evaluation quality that influence ratings on other factors. To assess the likelihood that specific team members “drive,” or have an impact on, other more reactive quality factors, MSI examined associations between these two types of team members. MSI also examined these relationships with the identification of external team leaders, the third team composition characteristic included in the meta-evaluation. What this analysis showed was that the presence of an evaluation specialist is statistically associated with several other quality factors, but the presence of local team members was not.

The quality factors that are directly associated with the presence of an evaluation specialist at a statistically significant level are shown below. All factors are statistically significant at the .05 level, and the first two listed are also significant at the .01 level:

- Annex included data collection instruments (#14)
- Report structured to respond to questions (not issues) (#17)
- Data collection methods linked to questions (#7)
- Data analysis method described (#10)
- Findings distinct from conclusions/recommendations (#23)
- SOW is included as a report annex (#33)
- Report indicated conflict-of-interest forms were signed (#15)
- Data analysis methods linked to questions (#11)

To illustrate the relationship between overall evaluation scores, which ranged from 0 to 10, and the presence of an evaluation specialist on an evaluation team, scores for evaluations that had an evaluation specialist were displayed in a scatter diagram and a line was fitted to these data. Figure 10 shows the relationship that emerges from this analysis. As the figure shows, the correlation between scores and the presence of an evaluation specialist is not perfect, yet the relationship is strongly positive.



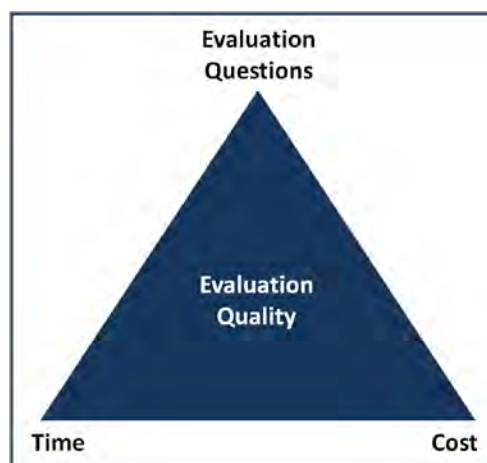
As a final step in this examination of associations between types of evaluation team members and other evaluation quality factors, MSI examined the extent to which the presence of an external team leader was associated with other evaluation quality score factors. This was done despite the absence of a strong association between the involvement of an external team leader and overall evaluation scores. This analysis showed that the presence of an external team leader is associated with several evaluation

quality factors, as listed below. These associations are statistically significant at the .05 level. Two of the quality factors associated with identifying an evaluation team leader differ from those associated with the presence of an evaluation specialist. Both, however, are associated with completion of USAID conflict-of-interest forms.

- Recommendations—clearly supported by findings (#32)
- Evaluation findings sex disaggregated at all levels (#27)
- Report indicated conflict-of-interest forms were signed (#15)

In addition to looking at the relationships between types of team members and evaluation quality factors, the meta-evaluation planned to examine the relationship between overall quality scores and three factors that have been mentioned in other meta-evaluations as possible “drivers” of evaluation quality: evaluation cost, duration, and the number of evaluation questions. As explained above, the cost and duration data needed for these analyses could not be obtained. MSI did, however, examine the relationship between the number of evaluation questions and overall evaluation quality.

This analysis did not find a statistically significant relationship between the number of evaluation questions and overall evaluation scores. The fact that different numbers of evaluation questions did not affect average scores runs counter to impressions held by many evaluators and some USAID staff. This is supported by comments from the meta-evaluation’s group interviews, the team leader survey, and previous studies on Agency evaluation practices.* This meta-evaluation’s finding that the number of evaluation questions is not statistically associated with overall evaluation quality suggests that while a large set of evaluation questions—more than 10 or even more than 20—may theoretically impede evaluation quality, teams are finding ways to deal with however many evaluation questions they are asked to address.



3. CONCLUSIONS

MSI’s mandate in the meta-evaluation was to determine whether evaluation quality had changed over time and to inform USAID about the aspects of evaluation quality on which the Agency excelled and where further work is needed. MSI was not asked to pinpoint the specific cause of changes in evaluation quality it observed. However, MSI would be remiss if it failed to recognize USAID’s multifaceted and consistent efforts to improve evaluation quality. From January 2010 onward, USAID efforts paralleled significant improvements in both the volume and quality of USAID evaluations identified by this study, most of which occurred between 2010 and 2012. Under the leadership of USAID Administrator Dr. Rajiv Shah, the Agency increased attention to, and investments in, evaluation as a key element in an effective development management system. Evaluation quality improvements measured by this meta-evaluation clearly correspond to the evaluation improvement initiative timeline shown in Figure 1, on page 2 of this report, which includes key dates associated with USAID’s promulgation of quality

*See in particular Blue & Clapp–Wincek, 2009; Hageboeck, 2009; Frumkin & Kearney, 2010—in the meta-evaluation bibliography.

improvements through speeches, new policy, intensive training, and supportive guidance materials, such as USAID’s How-To Note on Preparing Evaluation Reports.

At the beginning of this report, MSI described evaluation quality as the result of a partnership between evaluation clients, or evaluation managers, and the evaluation teams that carry out evaluations. A diagram with overlapping circles, in Figure 3, highlighted USAID’s responsibility for a) the questions teams address, b) ideas about best methods, c) team composition requirements, and d) reporting standards. The other side of the diagram indicated that evaluation teams assume responsibility for a) elaborating design and methods ideas, b) collecting and analyzing data, and c) setting forth findings, conclusions, and actionable recommendations in a clear and logical manner. When organized in terms of an evaluation partnership framework, the meta-evaluation findings lead MSI to slightly different conclusions about impediments to evaluation quality and courses of action regarding each dimension of the partnership model.

USAID–Determined Evaluation Partnership Elements

- On USAID’s side of the evaluation partnership, performance—with respect to evaluation quality factors—appears to have been strong and to have improved over the study period, with one notable exception in the team composition area: evaluation specialists.
- A full understanding of the USAID side of the evaluation partnership and the quality implications of USAID decisions would benefit from data about evaluation time and cost, which USAID once collected but no longer systematically gathers and maintains.

Evaluation quality decisions primarily determined by USAID have yielded good and fair ratings on five out of 10 quality factors, as reprised in Table 11.

Table 11. Recapitulation of Quality Factors That USAID Determines

Evaluation Report Component	Specific Evaluation Quality Factor	2012 Quality Factor Rating Baseline
USAID Primarily Responsible		
Background on Program/Project	Background included	Good
Theory of Change	Theory of Change included	Fair
Purposes	Management purpose included	Good
	Questions linked to purpose	Good
Questions—Number	10 or fewer questions	Marginal
Team Composition	External team leader identified	Good
	Evaluation specialist identified	Weak
	Local team members identified	Marginal—May not always apply
Evaluation Cost and Duration	Duration of evaluation reported	No data
	Cost of evaluation reported	No data

Despite longstanding guidance requiring the presence of an evaluation specialist on evaluation teams, ensuring that every evaluation team includes this type of team member, appears, from meta-evaluation findings, to be an Achilles heel in USAID’s strategy for improving evaluation quality. The issue is more complex than is suggested by the fact that only 19 percent of USAID evaluations indicated that an evaluation specialist was part of the team. What makes this team composition gap a critical weakness is the fact that the presence of an evaluation specialist on a team is statistically associated with positive performance on several other important evaluation quality factors. In other words, as the meta-evaluation demonstrated under Question 3, if you put an evaluation specialist on an evaluation team, this

step toward professionalizing the evaluation process is likely to improve ratings on up to seven other evaluation quality factors.

Few Teams Report Including Evaluation Specialists

Achieving compliance with USAID guidance on including an evaluation specialist appears, from group interviews and the Team Leader Perceptions Survey, to be impeded by

- The high value USAID staff place on having evaluation teams made up of sector specialists, and concomitantly low value placed on ensuring that a skilled evaluator is a member of every evaluation team.
- The lack of a common understanding of what the term “evaluation specialist” means and the skill set USAID intends to have represented by having this type of person on each evaluation team.
- Uneven reporting in evaluations on the composition of the evaluation teams that conduct evaluations, which forces USAID to make inferences from scattered references about the presence of various types of required/recommended evaluation team members (external team leader, evaluation specialist, local team members).

Poor reporting on evaluation team composition makes it difficult to understand the extent to which USAID involves local professionals on its evaluation teams and, as a byproduct, strengthens local evaluation capacity.

The Number of Evaluation Questions and Evaluation Quality

USAID received a “marginal” performance rating on its goal of limiting the number of evaluation questions to 10 or fewer. Had the bar been set at five or fewer questions, based on USAID’s How-To Note on evaluation SOW preparation, it would not have fared any better. MSI tentatively concludes, based on the meta-evaluation findings, that

- Including more than 10 evaluation questions in an evaluation SOW does not preclude the possibility that an evaluation will achieve a high overall evaluation quality score.

This conclusion remains tentative because of the absence of data on evaluation cost and time, which MSI, along with other meta-evaluation study authors, has hypothesized interact with the number of evaluation questions to influence evaluation quality.

Shared Decision Elements of an Evaluation Partnership

Methodology is the evaluation quality dimension to which USAID and evaluation teams both contribute. Evaluation SOWs present USAID’s ideas, while evaluation teams provide theirs through proposals and feedback on the SOW. USAID staff, organizations that manage evaluations for USAID, and team leaders for recent evaluations describe a continuum of involvement on the client side. Such a continuum runs the gamut from fully specifying a methodology, to asking teams to propose methods, to processes that blend client and team ideas. The continuum described suggests that, rather than being a shared collaborative space, the overlap shown in the circles in Figure 3 above may lack structure, which may in turn explain why so few factors for which responsibility is shared earned “good” quality standard compliance ratings, as seen in Table 12.

Table 12. Recapitulation of Quality Factors Determined Jointly by USAID and Teams

Evaluation Report Component	Specific Evaluation Quality Factor	2012 Quality Factor Rating Baseline
Shared USAID and Evaluation Team Responsibility for Methods		
Evaluation Methods	Data collection described	Good
	Data collection methods explained question by question	Weak—New
	Data analysis described	Marginal
	Data analysis methods explained question by question	Weak—New
Study Limitations	Study limitations provided	Fair
Gender	Data by sex reported for all results levels	Weak
	Reporting on gender differential access/participation/benefits	Marginal
Broad Evaluative Focus	Unplanned results	Weak—May not always apply
	Alternative causes	Weak—May not always apply

Of the three components of the evaluation partnership model, this “middle ground” is the weakest for USAID in terms of producing high-quality evaluations. This was determined not only in terms of the percentage of evaluation quality factors that fell into weak and marginal clusters based on their 2012 ratings, but also through MSI’s in-depth examination of these factors. Such analyses resulted in slightly different conclusions in each of the methodology subareas:

- USAID had a 96 percent compliance rate with its expectation that data collection methods will be included in evaluation reports, and roughly 30 percent compliance with expectations for the same type of presentation on data analysis methods. Based on this, MSI concludes that USAID staff, team leaders, and the organizations that undertake evaluations for USAID are unaware of these expectations. This is regardless of the fact that these expectations were stipulated in guidance available over the life of the meta-evaluation. Additionally, matrix approaches for complying with these quality factor standards have been used in USAID-funded evaluation training programs for staff throughout the years covered by the meta-evaluation. The gap appears to be in lack of awareness of expectations rather than in a lack of capacity to comply.

Weak coverage of two other types of findings is also discussed under methods. The two types of findings are included here since any improvements in the way that reports cover these issues would depend on whether and how evaluation teams gather the data that would be required.

- **Gender.** The levels of results for which USAID evaluations did or did not collect sex-disaggregated data, or discussed gender differences, suggest to MSI that evaluation teams may not always understand what USAID is seeking in regard to gender, or how to obtain those kinds of data. There seemed to be an awareness of USAID’s interest in gender, but that awareness was not generally matched to methods that would have obtained data on higher-level results, such as technology adopted by men versus women, or revealed differences between men and women’s access to or participation in project interventions.
- **Unplanned Results and Alternative Causes.** Weak performance on the coverage in USAID evaluations of broad issues such as unplanned results of programs, or alternative causes for results, is less likely to be the result of a skills gap among evaluators than a combination of the fact that few evaluations include evaluation specialists on teams and the absence of requests for this type of information in evaluation SOWs.

Evaluation Team Decision Elements of an Evaluation Partnership

Beyond evaluation methodology, which is covered above as a shared responsibility, evaluation team decisions affect how they structure evaluation reports within the latitude provided by USAID guidance and the reporting section of its evaluation SOW. Quality aspects reflecting the way in which the team handles evaluation findings, conclusions, recommendations and the study's executive summary are also controlled largely by evaluation teams. Meta-evaluation ratings indicate that evaluation team performance is marginal in some of these critical areas. In addition, evaluation reports were rated as being weak with respect to a number of evaluation report annexes, though some of these involve requirements that were introduced during the years covered by the meta-evaluation and therefore should be viewed with a less critical eye.

Weak ratings in areas for which evaluation teams are largely responsible were offset, MSI notes, by strong improvements on several quality rating factors for which teams are responsible. These include eight of the 11 evaluation quality factors for which MSI reported improvements of more than 10 percentage points between 2009 and 2010, as shown in Table 13.

Table 13. Recapitulation of Quality Factors Determined by Evaluation Teams

Evaluation Report Component	Specific Evaluation Quality Factor	2012 Quality Factor Rating Baseline
Evaluation Team Primarily Responsible		
Report Structure	Structured to answer questions	Fair
	Questions same as in SOW	Fair
	Questions addressed in report (not annexes)	Fair
	Written approval for changes in questions obtained	Weak—New
	Reason provided if some questions were not addressed	Weak—May not always apply
Findings	Social science methods evident in presentation of findings	Good
	Findings supported by data from range of methods	Good
	Findings distinct from conclusions/recommendations	Marginal
	Findings are precise (not simply “some, many, most”)	Fair
Recommendations	Recommendations—not full of findings, repetition	Fair
	Recommendations—specific about what is to be done	Fair
	Recommendations—specify who should take action	Marginal
	Recommendations—clearly supported by findings	Fair
Annexes	SOW is included as a report annex	Fair
	Annex included list of sources	Good
	Annex included data collection instruments	Good
	Report indicated conflict-of-interest forms were signed	Weak—New standard
	Statements of differences included as an annex	Weak—May not always apply
	Report explains how data will transfer to USAID	Weak—New standard
	Evaluation SOW includes Evaluation Policy Appendix I	Weak—New standard
Executive Summary	Executive summary mirrors critical report elements	Marginal

Based on the meta-evaluation’s findings on factors largely controlled by evaluation teams, MSI concluded the following:

- Factors that were rated as being weak may not represent a long-term problem. All of these factors involved standards that were either recently introduced or apply only in special situations and would thus not be expected to rise in the manner expected when a rating of 100 percent would be desirable.
- Three evaluation quality factors that received ratings placing them in the marginal cluster are much more problematic since, in principle, 100 percent of USAID evaluations should do these things well.

Substandard Quality on Evaluation Basics

Fewer than 50 percent of USAID evaluations between 2009 and 2012 complied with basic “good practice” in evaluation with respect to

- (a) Separating findings, conclusions, and recommendations
- (b) Specifying who should act on recommendations
and
- (c) Preparing executive summaries that accurately mirror the contents of an evaluation report

The fact that all three of the “marginal” evaluation quality factors are aspects of basic evaluation practice leads MSI not to a new conclusion but rather back to the finding that 80 percent of USAID evaluation teams either did not include an evaluation specialist or, if they did, did not report that such an individual was a member of their team.

While the ratings discussed here are on the evaluation team side of an evaluation partnership, USAID, as the client for evaluations, has an ongoing responsibility for quality control over these products. Small-group interviews and responses from team leaders of recent USAID evaluations both described evaluation quality control steps they were aware of or had been asked to participate in. These included steps both before and after fieldwork. Pre-fieldwork steps included reviews and approval of evaluation designs and methods, including instruments and sampling plans. Post-fieldwork steps included analysis briefings on findings, conclusions, and recommendations before the preparation of draft reports, and reviews of draft and final evaluation reports using a checklist that is available on USAID’s website. All of the quality control steps and tools mentioned by interview group participants and survey respondents are well suited for addressing quality ratings noted as being marginal in Table 13. This suggests to MSI that:

- Evaluation quality control procedures and tools used by some USAID staff are either not as widely known or as widely used throughout the Agency as might be needed to reach USAID’s evaluation quality goals.
- If the staff at USAID who are directly responsible for reviewing, commenting on, and accepting evaluation reports applied the systematic use of a simple evaluation quality factor checklist similar to the one used in this meta-evaluation, they could eliminate most of the weak, marginal, and even fair ratings in this final cluster, where evaluation teams have primary responsibility for the work, but USAID retains responsibility for oversight and quality of the products it procures.

Through the review of the elements in USAID’s evaluation quality partnership, MSI concluded that, in regard to performance and impediments to improved evaluation quality in each of the partnership model’s domains:

- At least one serious evaluation quality issue in each of the partnership model domains warrants attention.
- Given the many other demands on USAID staff’s time and resources, an ad hoc approach—trying a little bit of everything—to address weak and marginal aspects of evaluation quality would likely be less effective than prioritizing action on a few high-leverage interventions.
- Meta-evaluation findings that pinpoint specific actors in the evaluation process as being associated with, or having opportunities to affect, ratings on multiple evaluation quality factors

suggest that it may be possible to make substantial gains in USAID evaluation quality scores in the future by making just a few changes now.

- Assigning a high priority to improving weak evaluation quality ratings on two factors that aim to enhance USAID’s understanding of how programs and projects affect men and women will not dramatically raise USAID’s overall evaluation quality scores, but it could help the agency achieve policy goals in that arena, and may thus simply be the “right thing to do.”

4. RECOMMENDATIONS

The meta-evaluation of USAID 2009–12 evaluations presented in this volume found the quality of evaluation reports increased over the study period. It also identified evaluation quality factors where practice does not yet fully reflect USAID’s evaluation standards. As USAID is already engaged in a multiyear initiative aimed at increasing evaluation quality and effective utilization of USAID evaluations to support evidence-based decision-making throughout the program cycle, it might be assumed that the evaluation quality issues identified in this report will resolve themselves in the course of the Agency’s larger evaluation improvement effort. To a certain extent, that is likely to be the case, particularly with respect to improved compliance with evaluation requirements that were issued a year or two into the period covered by the meta-evaluation. Other types of weaknesses in USAID evaluation quality, some of which have been noted in previous meta-evaluations, may require direct attention to correct.

As the conceptual framework for this meta-evaluation suggested, evaluation quality is driven by decisions made, and actions taken, by two groups of people. The first are clients, who identify the need for evaluations, develop the SOWs that guide them, and receive and use evaluation products. The second are evaluation teams, who conduct these evaluations and deliver empirical findings and recommendations in response to client SOWs. USAID’s broad evaluation improvement initiative already focuses on policies, procedures, guidelines, and trainings that are intended to change USAID staff knowledge and practices, and through them the practices of organizations and individuals that undertake evaluations for the Agency. Recommendations from this meta-evaluation, to be helpful, must supplement rather than duplicate those efforts as efficiently as possible. With this in mind, MSI offers three recommendations to USAID/PPL/LER, which individually and collectively can help enhance the quality of USAID evaluations and evaluation reports in precisely those areas that offer opportunities for improvement. Of these three, the first recommendation is considered the most important for systematically raising the quality of evaluations across all sectors and regions.

Recommendation 1. Increase the percentage of USAID evaluations that have an evaluation specialist as a fulltime team member with defined responsibilities for ensuring that USAID evaluation report standards are met from roughly 20 percent as of 2012 to 80 percent or more.

Findings from the meta-evaluation demonstrate that the presence of an evaluation specialist on teams is associated, at high level of statistical significance, with improved ratings on a host of evaluation quality features that are a function of what evaluation teams do once provided with a USAID evaluation SOW. Implementing this recommendation may require an effort to change views among midlevel USAID managers regarding the types of skills needed on evaluation teams, but it will not require new policy or significant monetary resources. To radically increase the percentage of evaluation teams that include an evaluation specialist may, however, require senior staff support; increased awareness of USAID’s existing requirements among senior and midlevel managers and contracting officers; and possibly an upgrade in the language used in USAID guidance on this matter, such as a change from “should” to “must” or

MANDATORY. Making this single change is highly likely to yield measureable improvements in future meta-evaluation ratings on a variety of evaluation quality measures of the sort used in this study.

Lack of a common understanding in USAID of the knowledge and skills an “evaluation specialist” is expected to bring to bear on evaluation planning, implementation, and the preparation of reports impedes effective action on this recommendation. Accordingly, Exhibit A maps the evaluation knowledge and experience continuum to help USAID staff select appropriate individuals as “evaluation specialists” for evaluation teams.

As a corollary to this recommendation, enhance USAID’s description of the required elements of an evaluation report to include a section or requirement to describe the composition of the evaluation team, including the names of each evaluation’s team leader and evaluation specialist. This will allow USAID to track whether the percentage of evaluations with these team members, as well as local evaluation team members, is increasing. It also will make it easier for USAID to reach out to those individuals for feedback when desired. When modifying evaluation reporting guidance to capture information on the composition of evaluation teams, USAID could also address its lack of information on evaluation duration and costs, recognizing that any new requirement of this nature, like most performance indicators, will require standardized definitions of cost and duration.

Recommendation 2. Intervene with appropriate guidance, tools, and self-training materials to dramatically increase the effectiveness of existing USAID evaluation management and quality control processes.

MSI’s second recommendation is intended to complement its first recommendation. USAID staff are already responsible for ensuring that evaluations they oversee meet Agency quality standards. Findings from the meta-evaluation indicate that while they spend time on this function the energy USAID evaluation managers have invested over the years has not eliminated evaluation quality problems. As reported, some USAID staff are aware of and use a variety of evaluation quality control procedures and tools, but persistent evaluation report quality issues indicate this knowledge is not as widely shared and applied as would be desirable. The knowledge and tools needed to improve evaluation quality control at the evaluation manager level already exist, but they are not well codified or as simple and straightforward as will be needed to significantly increase their use.

MSI is tabling this particular recommendation at what it appears may be a particularly opportune moment. USAID is currently implementing a Mission Order Standardization Initiative that encourages field offices to improve their evaluation review processes consistent with ADS 203.3.1.8. This guideline provides PPL/LER with an important opportunity to link all missions to a small set of guidelines and tools for improving evaluation quality. Something as simple as making the use of an evaluation quality checklist, similar to the one used in this study, a standard operating procedure for USAID staff reviewing draft evaluation reports and accepting final versions has the potential for dramatically improving USAID evaluation quality factor ratings. This is particularly true for numerous new evaluation quality standards factors where evaluation teams have significant responsibilities. To help USAID implement this recommendation, Exhibit 2 summarizes existing USAID evaluation management and quality control practices that may warrant scaling up.

As a corollary to this recommendation, extend the improvement of the evaluation quality review process to include a more systematic approach for identifying and correcting weaknesses in evaluation SOWs in the field. Additionally, invest in sufficient research to determine the degree to which SOW quality problems are the source of, and/or exacerbate, evaluation report quality issues identified in this meta-evaluation.

Recommendation 3. As a special effort, in collaboration with USAID’s Office of Gender Equality and Women’s Empowerment, invest in the development of practitioner guidance materials specific to evaluation.

MSI’s final recommendation addresses a specific weakness identified by the meta-evaluation. While attention to gender considerations is expected to be a priority throughout the program cycle, the meta-evaluation identified evaluation performance on this issue as weak. Evidence from the study identified gaps in the sex-disaggregated data on results evaluations and relatively unsystematic efforts to document gender differential access, participation, and benefits in USAID programs and project. These weaknesses, in turn, suggest a lack of the understanding, skills, and tools needed to improve this aspect of the Agency’s evaluations. USAID’s investments in earlier years in the development of gender analysis tools for planning have served the Agency well. MSI recommends a parallel effort that builds on this base to create gender-specific evaluation techniques, tools, and guidance.

Exhibit 1: Evaluation Specialist Knowledge and Experience

Regardless of an individual's knowledge and experience of a sector (agriculture, health, education or other fields), an evaluation specialist has professional training that is relevant for the conduct of high quality evaluations as well as practical experience.*

	Evaluation Generalist		Evaluation Specialist		
	Novice	Journeyman	Novice	Journeyman	Master
Knowledge Dimension	<ul style="list-style-type: none"> 40 hour professional evaluation training program, <p>OR</p> <ul style="list-style-type: none"> Semester undergraduate course involving research design/methods, <p>OR</p>	<ul style="list-style-type: none"> 40 hour professional evaluation training program, <p>AND</p> <ul style="list-style-type: none"> Semester undergraduate course involving research design/methods, <p>OR</p>	<ul style="list-style-type: none"> 80 hour professional evaluation training program, <p>OR</p> <ul style="list-style-type: none"> Two or more undergraduate or graduate school courses covering research design/methods, <p>AND</p>	<ul style="list-style-type: none"> 80 hour professional evaluation training program, <p>AND</p> <ul style="list-style-type: none"> Two or more undergraduate or graduate school courses covering research design/methods, <p>AND</p>	<ul style="list-style-type: none"> Two or more undergraduate or graduate school courses covering research design/methods, <p>AND</p> <ul style="list-style-type: none"> Teaches evaluation courses or professional evaluation training programs, <p>AND</p>
Practice Dimension	<ul style="list-style-type: none"> Full member of one evaluation team involving field data collection, <p>OR</p> <ul style="list-style-type: none"> Full member of one evaluation design team that produced a design product 	<ul style="list-style-type: none"> Full member or Team Leader of one or more evaluation team involving field data collection, <p>OR</p> <ul style="list-style-type: none"> Full member or Team Leader of one or more evaluation design team that produced a design product 	<ul style="list-style-type: none"> Full member or Team Leader of one evaluation teams involving field data collection, <p>OR</p> <ul style="list-style-type: none"> Full member or Team Leader of one evaluation design team that produced a design product 	<ul style="list-style-type: none"> Full member or Team Leader of multiple evaluation teams involving field data collection, <p>OR</p> <ul style="list-style-type: none"> Full member or Team Leader of multiple evaluation design team that produced a design product 	<ul style="list-style-type: none"> Team Leader for multiple evaluations, <p>OR</p> <ul style="list-style-type: none"> Team Leader for multiple evaluation design that produced a product <p>AND</p> <ul style="list-style-type: none"> Technical quality oversight over a portfolio of evaluations

* There is no fixed academic curriculum through which evaluation specialists acquire knowledge, but, in general, the range knowledge expected covers both the design of evaluations, which is loosely based on standard research design principles, including hypothesis development and testing, as taught in multiple academic disciplines and social science methods for collecting and analyzing data. Methods subsumes techniques that are considered to be (a) quantitative (surveys of target populations based on probability samples), structured observation (checklists and other tools), and instruments (that measure weight, distance, substance quality and other quantifiable dimensions) analyzed using statistical or econometric techniques or (b) qualitative (semi-structured observation; group interviews and focus groups, informal data gathering from populations based on purposive samples and other methods for obtaining narrative or visual/auditory data analyzed using content or pattern analysis techniques or sometimes transformed into a format that allows quantitative analysis. Research report writing is normally an element of both quantitative and qualitative methods training.

Exhibit 2. Scaling-Up Use of “Good Practice” Quality Control Checkpoints

Evaluation SOW reviews and reviews of draft and final reports are well accepted practices in USAID, yet evidence from the meta-evaluation indicates that simple tools, such as already existing checklist, may not be used on a routine basis. If they were, their systematic use would have spotted many of the deficiencies noted in the meta-evaluation well before evaluation reports were finalized and approved. Further, both USAID staff and evaluation team leaders identified other evaluation quality checkpoints and tools with which they were familiar, but which do not appear to be widely used by all evaluation managers or Missions. Building on and scaling up such practices would likely help USAID improve its evaluation quality ratings.

Evaluation Quality Checkpoints	Timing	Tools	Product or Deliverable
Evaluation SOW Review	Prior to SOW approval	<ul style="list-style-type: none"> • SOW How-To Note (<i>already in place</i>) • SOW Checklist – targeted to issues that affect evaluation quality. (<i>Shortening currently published version and preparing a simple handbook to guide users are recommended</i>) • SOW “Good Examples” publication (<i>may warrant review and update to match updated ADS and How To</i>) 	<ul style="list-style-type: none"> • Final, approved SOW that complies with USAID standards
Evaluation Team’s Document Review (or Desk Study)	Prior to completion of Detailed Evaluation Design	<ul style="list-style-type: none"> • Template for Summarizing information found in documents by evaluation question and identifying information gaps that remain to be filled in order to answer evaluation questions • Consider a How-To Note 	<ul style="list-style-type: none"> • Document Review Report – standalone document or as 1st part of an inception report
Detailed Evaluation Design (<i>prepared by the team that will actually conduct the evaluation; supersedes proposal stage</i>)	Prior to approval to start evaluation field work/data collection (<i>precondition for utilization of LOE allocated for field work</i>)	<ul style="list-style-type: none"> • Illustrative evaluation design report outline • Sample “Getting to Answers” matrix for associating design choices and data collection/analysis methods to evaluation questions for inclusion in design report • Consider a How-To Note 	<ul style="list-style-type: none"> • Evaluation Design Report – standalone document or as 2nd part of an inception report.
Post Field Work & Analysis Review of Completeness of Findings, Conclusions, and Recommendations	Prior to approval for utilization of LOE allocated for writing F-C-R sections of a draft report	<ul style="list-style-type: none"> • Example of bulleted presentation of key points for an F-C-R review, illustrating: findings may support multiple conclusions and recommendations; recommendations may draw on multiple findings. • Review guide for this step - purpose is to spot unaddressed questions and gaps in evidence and/or logical exposition. • Consider a How-To Note 	<ul style="list-style-type: none"> • Instructions to team on remedial work required, as appropriate • Approval for drafting the evaluation report
Review of Draft Evaluation Report & Approval of Final	Prior to providing team with feedback on draft and prior to approval of final evaluation report	<ul style="list-style-type: none"> • Evaluation Report How-To Note and template (<i>already in place</i>) • Evaluation Report Review Checklist - targeted to issues that affect evaluation quality. (<i>Shortening currently published version & preparing a simple handbook to guide users are recommended</i>) 	<ul style="list-style-type: none"> • Feedback on draft report, prior to – • Approved Final Evaluation Report

PART 2. DETAILED META-EVALUATION FINDINGS

Part 2 of this report provides readers with in-depth information on the findings of USAID's 2009–12 meta-evaluation. It also provides basic characteristics of USAID evaluations and MSI's responses to three questions addressed by the meta-evaluation regarding the quality of USAID evaluations. Part 2 is best understood as a more detailed presentation of the findings from Part 1 of this report. The key differences between Part 2 and the findings section of Part 1 are the coverage of meta-evaluation findings from regional and sector perspectives, and comparisons of USAID Forward evaluations with other evaluations completed between July 2011 and December 2012. This treatment of the data complements all aspects of the meta-evaluation findings on an overall and annual basis. This section starts with a brief synopsis of the sources of evidence for this study.

OVERVIEW

This section provides detailed findings not presented in Part I. The evidence presented in Part I is drawn from multiple sources, including

- A review of the coverage of previous meta-evaluations as summarized in Annex C
- Data extracted from 340 evaluations using 1) a Basic Evaluation Characteristics Description Instrument (Annex C) to extract information about evaluation features determined by USAID and 2) an Evaluation Quality Factor Review Checklist (Annex C) largely by evaluation teams
- An Evaluation Team Leader Perceptions Survey (Annex D)
- Small group interviews with USAID technical and regional bureau staff and organizations that conduct evaluations (summarized in Annex E)

A full description of the study methodology is provided in Annex B.

For readers interested in accessing specific information covered in Part 2, data tables by region, sector, and USAID Forward evaluations are provided at the end of this section. They are presented with a chart that displays findings about each of the evaluation quality factors examined by this study, which applied analytic techniques to address the three evaluation questions the study attempted to answer.

I. BASIC EVALUATION CHARACTERISTICS IN DEPTH

The meta-evaluation presented in this volume is based on 340 USAID evaluations completed between January 2009 and December 2012 and submitted to the DEC, which categorizes and stores a wide range of USAID documents that are generally available to the public.

This section of the meta-evaluation provides an overview of several characteristics of USAID evaluations as well as evaluation management and quality control processes determined by USAID—often well in advance of selecting an evaluation team. Evaluation and evaluation process characteristics covered in this section include six of 10 “primarily client-determined” aspects of evaluation, which are identified in Part I of this report. The characteristics discussed include

- USAID region or bureau
- Sector or topic on which the program or project focused
- Scope of the evaluation—one project or program, or several
- Timing—during implementation, towards the end, or ex-post
- Type of evaluation—performance or impact, and the type of design if impact
- Evaluation budget or cost
- Duration of evaluation
- Evaluation SOWs as a determinant of quality
- USAID’s evaluation management and quality control process
- Designation as a USAID Forward evaluation

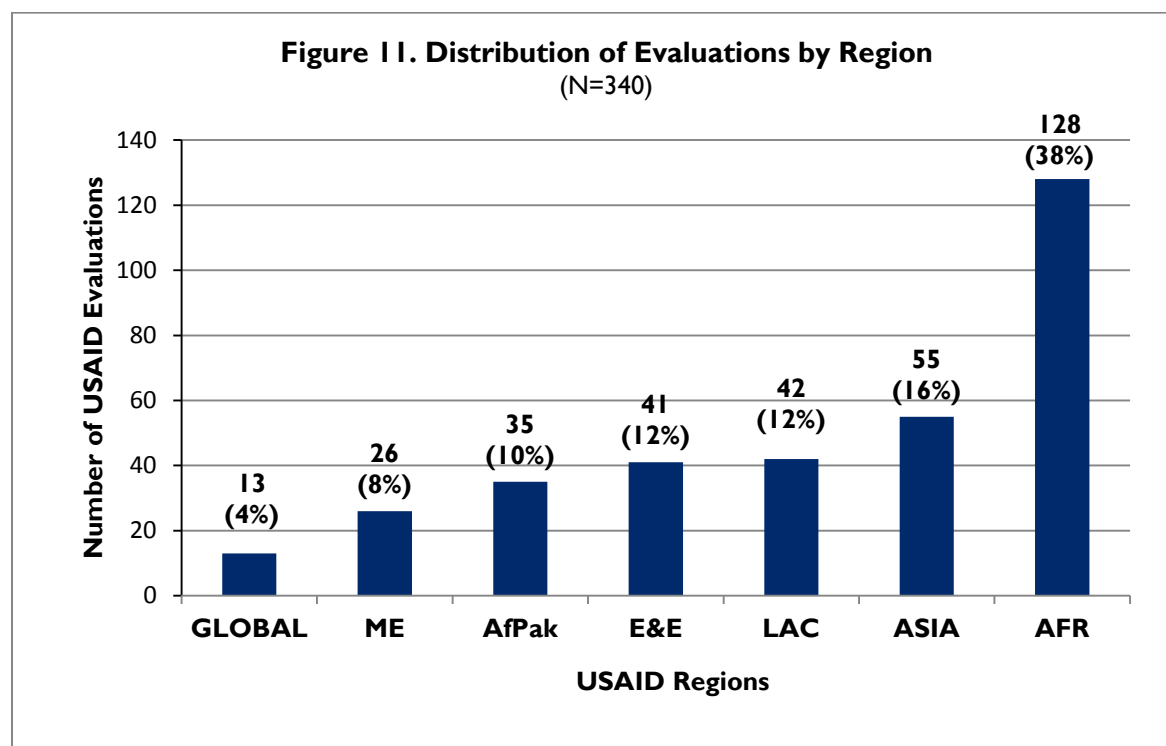
Each of these basic characteristics is discussed in this section before turning to the meta-evaluation’s findings on evaluation quality in the next section.

A. USAID Region or Bureau

Evaluations in this sample come from all regions in which USAID works as well as its technical bureaus in Washington D.C. which will be referred to as USAID/W or occasionally by the term “Global” in this study. The distribution of sampled evaluations from these geographic locations is shown in Figure 11 below.

As this figure shows, the largest share of evaluations in the sample comes from USAID’s Africa Bureau. Evaluations from Africa account for 38 percent of all evaluations included in the meta-evaluation. This figure is higher than in previous meta-evaluations, where evaluations from Africa accounted for roughly 25 percent of those examined. African countries represented by relatively large numbers of evaluations in this meta-evaluation include Ethiopia (16), Uganda (12), and Sudan/South Sudan (10). Large numbers of evaluations for individual countries in this meta-evaluation sample also include 35 evaluations from the Afghanistan–Pakistan (AfPak) region, most of which are from Afghanistan. Iraq is represented by 10 evaluations in the Middle East Cluster. Ten or fewer evaluations are included for all other individual countries represented in the meta-evaluation sample.

While the percentage of evaluations from Africa rose over previous meta-evaluations at the regional level, percentages of evaluations for other regions decreased. Asia, for example, decreased from 25 percent of the sample to 16 percent of evaluations reviewed, while Latin American and Caribbean (LAC) evaluations decreased from 12 percent in the last meta-evaluation to 8 percent in the current study and the Middle East (ME) dropped from 8 percent to 5 percent. Conversely, the percentage from Europe and Eurasia (E&E) rose from 11 percent to 12 percent and AfPak, represented by 35 evaluations, is a new region.



Within the geographic regions, numbers of evaluations vary by country. In Figures 12 through 17, the percentage of country-specific evaluations is shown, by region, for all countries where several evaluations were rated. Countries where only one or two evaluations were rated are grouped as “other” in these figures.

Figure 12. AfPak Region Evaluations
(N=35)

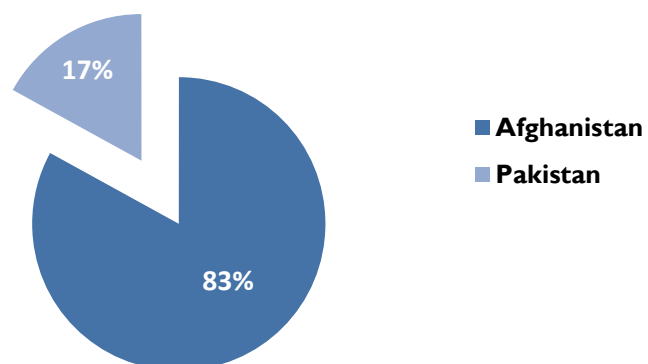


Figure 13. Africa Region Evaluations
(N=128)

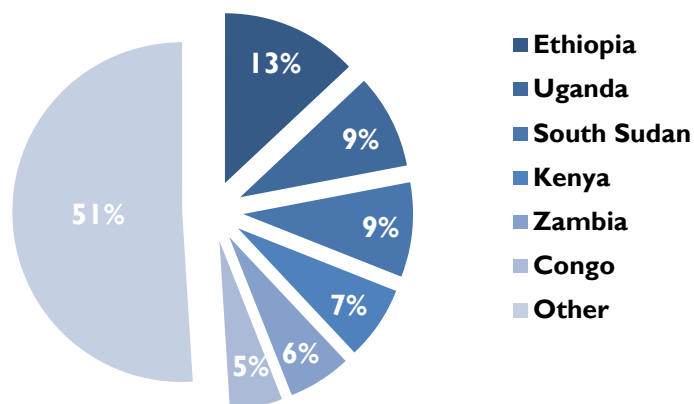
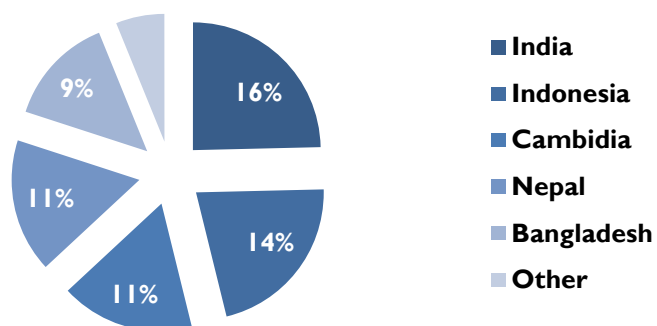
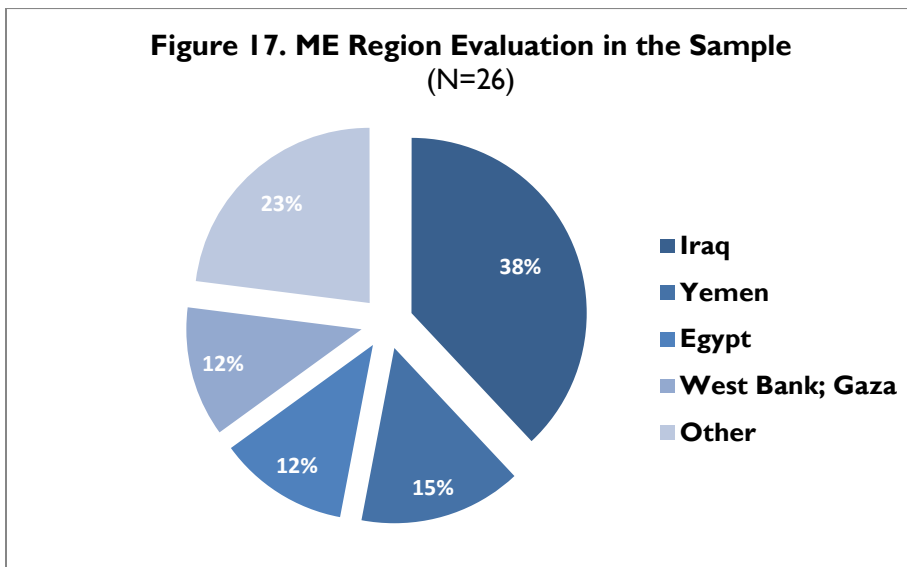
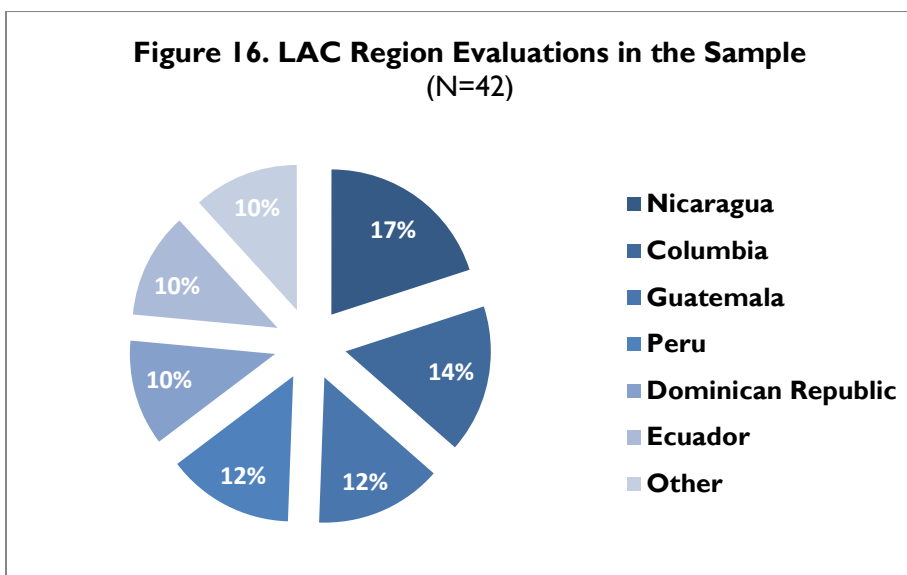
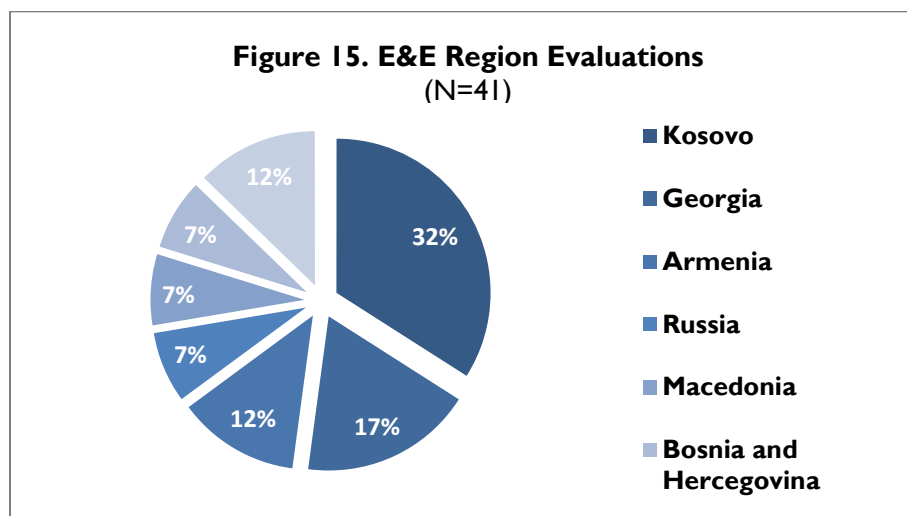


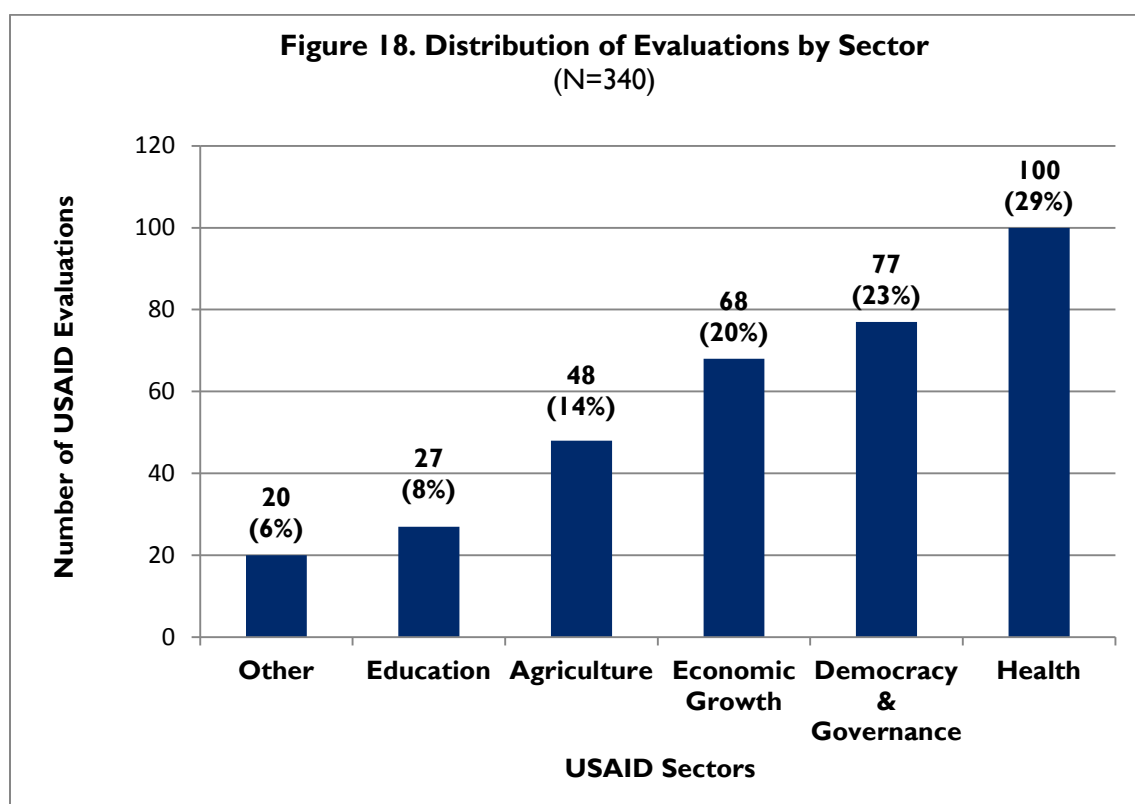
Figure 14. Asia Region Evaluations
(N=55)





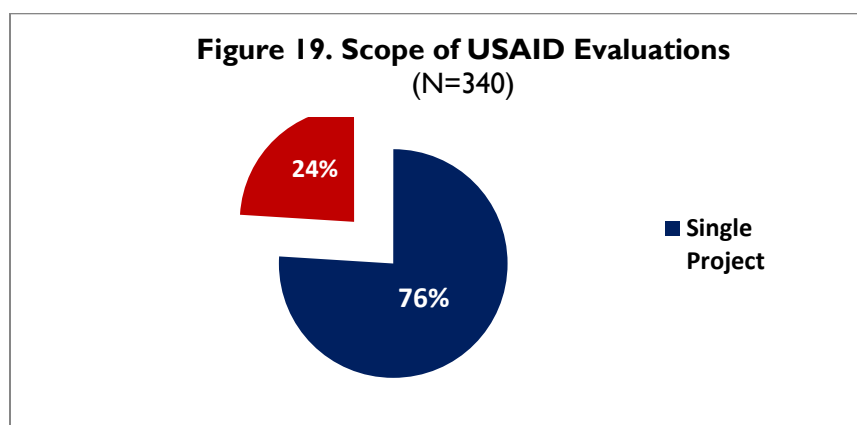
B. Sector or Topic of the Program or Project Evaluated

On a sector basis, evaluations of health projects and programs dominate the meta-evaluation sample as shown in Figure 18 below. Notably, the representation of health projects in the meta-evaluation sample at 29 percent is lower than this sector's representation in the 2005–08 meta-evaluation, where health evaluations accounted for 38 percent of the sample. Of the health evaluations included in the 2009–12 sample, 58 percent were undertaken in Africa. Other shifts between the current meta-evaluation and previous studies include an increase in the percentage of Democracy and Governance (DG) evaluations from 13 percent to 23 percent, with a large share of these (26 percent) coming from the E&E region. A larger share of evaluations in the sample was also found for Economic Growth (EG) projects and programs, which rose from 11 percent to 20 percent and includes a relatively large group of evaluations from Asia (26 percent of all EG evaluations in the sample). Shares of the meta-evaluation sample for two other sectors—agriculture/natural resource management and education—remained roughly the same as they were in the 2005–08 meta-evaluation.

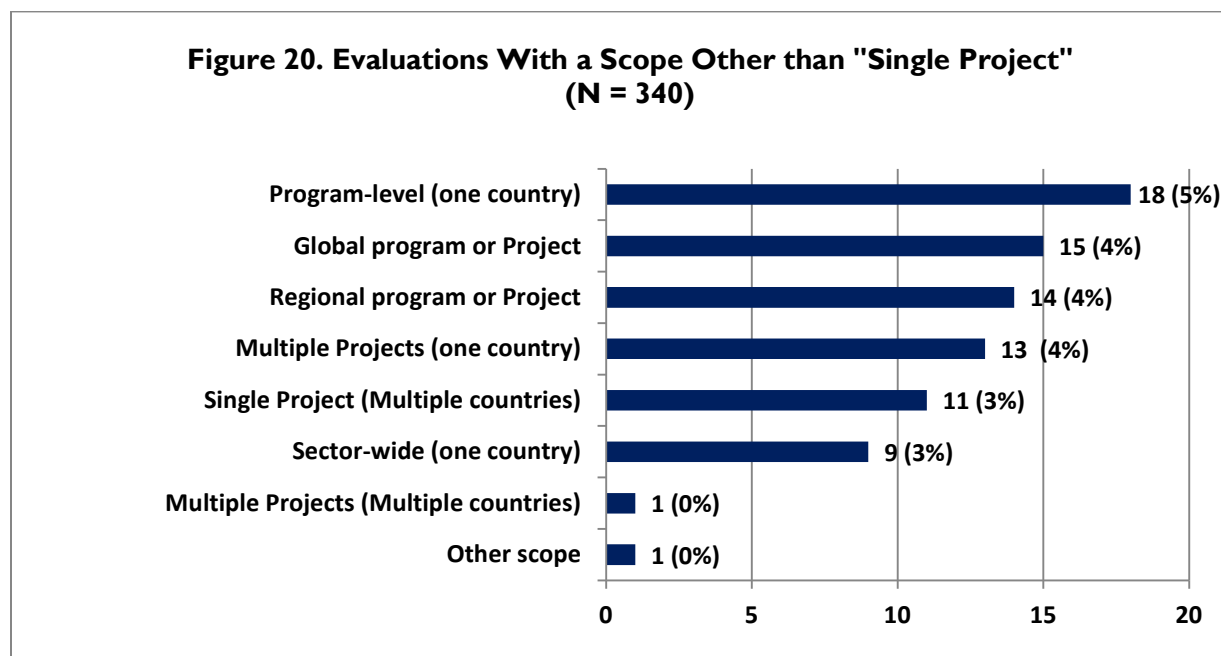


C. Scope or Scale of the Evaluation

As in previous meta-evaluations, the distribution of this current meta-evaluation indicates that the primary focus of USAID evaluations is individual projects. Of the 340 evaluations examined by this study, 258 (76 percent) focused on a single project as illustrated in Figure 19. This percentage parallels a finding from the previous (2005–08) meta-evaluation in which 74 percent of all evaluations focused on individual projects. This finding is consistent across USAID regions and sectors.



The 82 evaluations in the sample of 340 (24 percent) that did not focus on a single project were scored as having a variety of other characteristics. Some of these 82 evaluations were categorized according to more than one option; for example, the rating system allowed them to be classified as both multiproject and regional. Accordingly, Figure 20 below is illustrative of other focuses in USAID evaluations, and does not provide an exact breakdown.

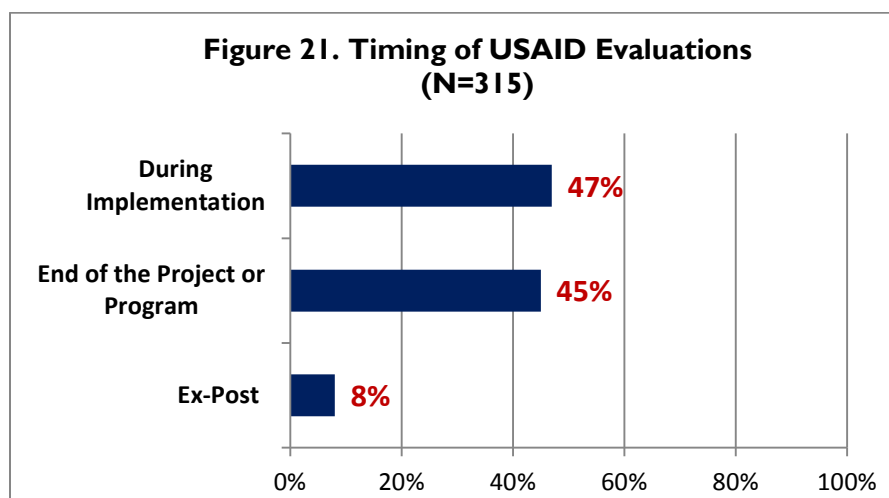


D. Evaluation Timing

For 315 (93 percent) of the 340 evaluations reviewed, the timing of USAID evaluations were clear from evaluation reports. Among those for which timing decisions were clear, there were roughly equal divisions between evaluations undertaken during project or program implementation and those undertaken towards or at the end of a program or project, as shown in Figure 21. These figures are roughly equivalent to those for the 2005–08 meta-evaluations, which categorized 46 percent of the evaluations it rated as formative, which is roughly consistent with “during implementation,” and 43 percent as summative.

In addition to these timing choices, MSI found 26 evaluations which are best classified as being ex-post, meaning that the evaluations were deliberately planned to be conducted after the project or program

had ended. Evaluations in this cluster were sometimes undertaken on an ex-post basis because funds had not been available earlier, or because USAID recognized that it could still benefit from lessons that evaluations might yield after a program or project has ended.



MSI's analysis of evaluation timing by year, region, and sector showed that for the most part there was a fairly evenly distribution across dimensions within each group. Further analysis showed that other factors over which USAID has control, such as the identification of a management purpose for an evaluation and the number of questions that teams are expected to address, were rated as being roughly equivalent for evaluations conducted during implementation and for those carried out closer to the end of a program or project.

E. Management Purpose

Closely related to evaluation timing is the management purpose of an evaluation, which USAID provides in an evaluation SOW and which evaluation teams are expected to reflect upon when writing their reports.

Information extracted from evaluations reviewed by the meta-evaluation team indicate that for the 2009–12 period, the most frequently cited management reasons for undertaking evaluations were to a) improve the implementation or performance of an ongoing project or program (41 percent of evaluations), b) facilitate the design of an immediate follow-on, for example, a Phase 2 project (15 percent), or c) provide input or lessons to support the design of future strategies, programs, or projects not directly related to the program or project being evaluated, for example, in a different country or sector (46 percent).

MSI's review of earlier USAID meta-evaluations indicate that the main reasons for undertaking evaluations in 2009–12 were also frequently cited as management purposes in earlier periods, as Table 14 indicates.

Table 14. Historical Data on Evaluation Purposes

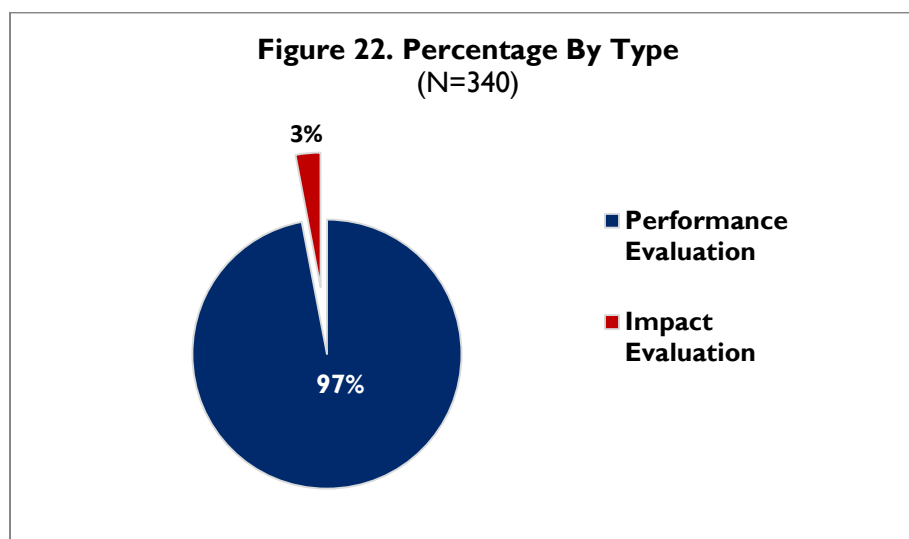
Evaluation Purposes				
Year	Improve Implementation/ Decisions About Current Program/Project	Support Design of an Immediate Follow-On (e.g., Phase 2)	Facilitate Design of Future Strategies/ Programs/Projects	Other
1989–90	22%	21%	22%	35%
1998–99	22%	37%	14%	27%
2005–08	46%	15%	17%	22%
2009–12	43%	15%	44%	4%

F. Type of Evaluation

USAID's Evaluation Policy updated USAID's typology of evaluations characteristically undertaken by the Agency when it listed the two types of evaluations it intends to undertake going forward, namely:

- **Impact evaluations**, which measures the change in a development outcome that is attributable to a defined intervention using a rigorously defined counterfactual to control for factors other than the intervention that might account for the observed change. In practice, such evaluations involve comparisons between beneficiaries of an intervention and a comparison group that did not receive the intervention.
- **Performance evaluations**, which examines what a particular project or program has achieved, often through before-and-after comparisons; how results were or are being achieved; how the project or program is perceived and whether it is valued; and other questions for which USAID managers need answers.

Applying these definitions, the meta-evaluation team identified 11 impact evaluations (3 percent), as shown in Figure 22. All of the evaluations coded as being impact evaluations met two criteria: 1) they included a comparison group that satisfied the requirement for a counterfactual and 2) they each had data from at least two points in time. Of these 11 impact evaluations, 10 employed quasi-experimental designs and one used an experimental design with randomized assignment.



On page 41 of its 1970 Evaluation Handbook, USAID carefully described the “ideal type” of evaluation design for examining causality by making comparisons, which identified randomized assignment as the best way to establish control groups for this purpose. Given the existence of this guidance, and its similarity to USAID’s 2011 Evaluation Policy, and the preference for experimental designs over quasi-experimental designs for impact evaluations, MSI reviewed prior meta-evaluations to discern the extent to which USAID evaluations included control or comparison groups of the type now expected for USAID impact evaluations. One data point on this question came from page 34 of USAID’s 1987–1989 meta-evaluation where it stated that among “evaluations using various analytical methods, 12% of the 287 evaluations rated made some use of comparison or control groups.” This rating factor appears to have included the unanticipated use of existing data to construct such comparisons.

While classifying evaluations for the meta-evaluation, MSI noted that the distinction between performance and impact evaluations is not as precise in practice as it is on paper. Some of the evaluations MSI reviewed that used the term “impact evaluation” focused not only on questions about causality and attribution but also on performance questions, as defined by USAID. Among the 340 evaluations that made up the meta-evaluation sample, 94 (28 percent) identified questions about causality or attribution as part of their mandate. Regionally, questions about causality were more frequently asked in Asia and E&E evaluations than elsewhere, and were particularly low in LAC and Global evaluations. On a sector basis, agriculture/natural resources management and DG evaluations included more questions about causality than did health evaluations, which were least likely to include specific questions of this sort.

With only 11 true impact evaluations, it could be inferred that the remaining 83 evaluations used non-experimental methods to answer these types of questions. MSI’s review of evaluation methods did not, however, identify a large number of well-articulated and well-implemented non-experimental designs. This classification experience paralleled some of the comments offered by USAID staff and firms that conduct USAID evaluations during small group interviews.

In a small group interview with USAID technical bureau staff, one participant noted that many people at USAID still have trouble differentiating between performance and impact evaluations. Comments from one regional bureau representative and four representatives of firms converged in reporting that there is a lot of confusion at USAID as to when an impact evaluation is appropriate. A representative of one of the firms added that not all impact evaluations are actually real or high-quality impact evaluations. In addition, an evaluation firm representative stated that the people writing evaluation requests for proposals (RFPs) want to see impact, but do not understand the methodological and budgetary implications of an impact evaluation, while another added that the RFPs for impact evaluations are too vague to be accurately responded to. On a related point, representatives of four firms indicated that many recent SOWs ask for both performance and impact questions in the same RFP.

Commenting on how USAID might guide its impact evaluation practices, one technical representative stated that it was very difficult to get anyone to undertake an impact evaluation previously, but now it is much more common. A representative of one of the firms added that USAID needs to better incorporate impact into project designs, while a representative from another firm said that while USAID is looking to have projects and impact evaluations start simultaneously, USAID procurement policies prevent this from happening. Three other firms disagreed, indicating that they have seen projects and impact evaluations undertaken concurrently. Elaborating on this situation, another firm commented on how project implementers do not always agree to the conditions necessary to conduct an impact evaluation involving randomized assignment, further complicating the ability to them.

In addition to the comments above, and a cautionary tale about correctly coding impact and performance evaluations from the meta-evaluation team’s experience, MSI’s rating instrument also

collected data on the presence of causality questions. As it turned out, a larger number of evaluations included these questions than could be properly counted as an impact evaluation.

Not all evaluation reports included evaluation questions, as will be further discussed in the report. Out of 215 evaluations that included or referenced the existence of questions, 94 (44 percent) included questions about causality. This number far exceeds the 11 evaluations coded as impact evaluations or hybrids; it suggests that questions about causality are also embedded in roughly 40 percent of evaluations in the performance evaluation cluster. When disaggregated by year, study data showed that the percentage of evaluations that include questions about causality was reasonably consistent over time.

Notably, with regard to evaluation questions and causality, is the fact that in some cases—mostly impact evaluations but also in some performance evaluations that focus on causality—no formal questions were provided in the description of the evaluation design. Instead, the design only described the intervention and the outcome (or dependent) variables to be measured. Evaluations with this type of structure, which is both rigorous and appropriate for formally structured impact evaluations, may have lost a point in the evaluation rating for not including a list of questions similar to those found in most performance evaluations.

G. Evaluation Questions

When developing evaluation SOWs, USAID staff determine the questions they expect an evaluation team to address. As early as 2003, USAID ADS 203.3.1.4 encouraged Agency staff involved in planning evaluations to identify a “small number” of key questions an evaluation should address and to include those question in an evaluation SOW. Over the years, a consensus emerged within the Agency that a “small number” meant 10 or fewer questions.* With that tacit agreement in mind, MSI counted the number of evaluation questions raised in evaluation reports between 2009 and 2012 or in the evaluation SOWs attached to those reports, and then clustered them into three groups:

- 1 to 10 questions, which it treated as the USAID quality standard for this meta-evaluation
- 11 to 20 questions, as a second cluster
- 21 or more questions, to identify evaluations with a very large number of questions

In addition to establishing these clusters of number of questions, MSI identified and applied three distinct ways of counting questions:

- The count of numbered questions in a SOW or evaluation report, which easily captures numbered questions but may skip over sub-questions or questions that are not numbered.
- The count of visible question marks in a set of questions. This method captures all sub-questions as well as instances in which questions are presented as bullets or in some other unnumbered manner.
- The count of explicit and implicit question marks in a set of evaluation questions.

Each of these methods yield different answers; for example, in the 133 evaluations that actually numbered their evaluation questions, the average number of “numbered questions” averaged 12 questions per evaluation. When counting the explicit number of question marks, as opposed

Impact Evaluations

“Impact Evaluations are being driven by a notion of compliance and being done not necessarily because they are promising, but because they feel they have to.”

Representative of a Firm
Conducting Evaluations for USAID

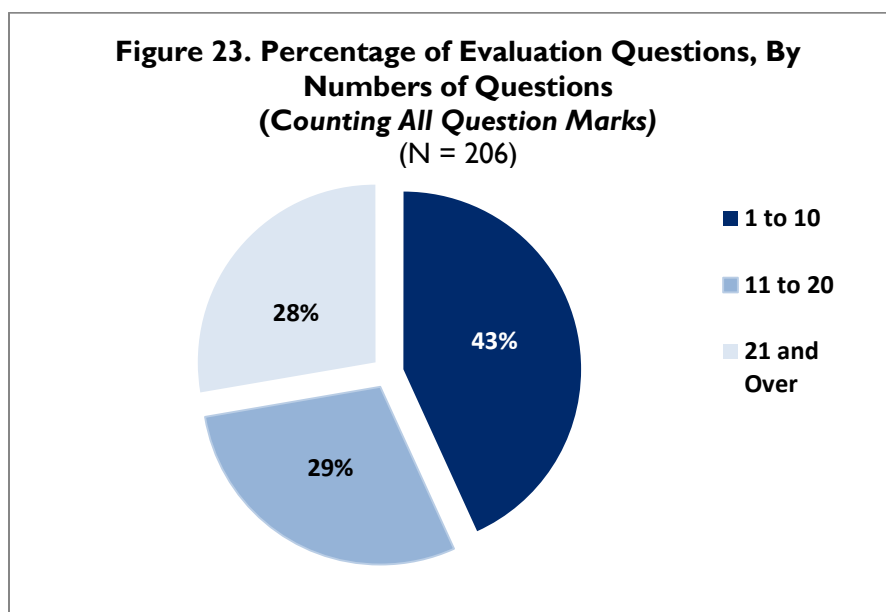
*More specifically, MSI’s decision to use the number 10 to represent the upper limit of USAID’s “small number” was based on the response of USAID participants, in USAID-funded evaluation trainings conducted over the past decade, when asked how they interpreted the ADS on this point.

to numbered questions, the average number of questions per evaluation rises to 19. Including implicit question marks would drive the count still higher.

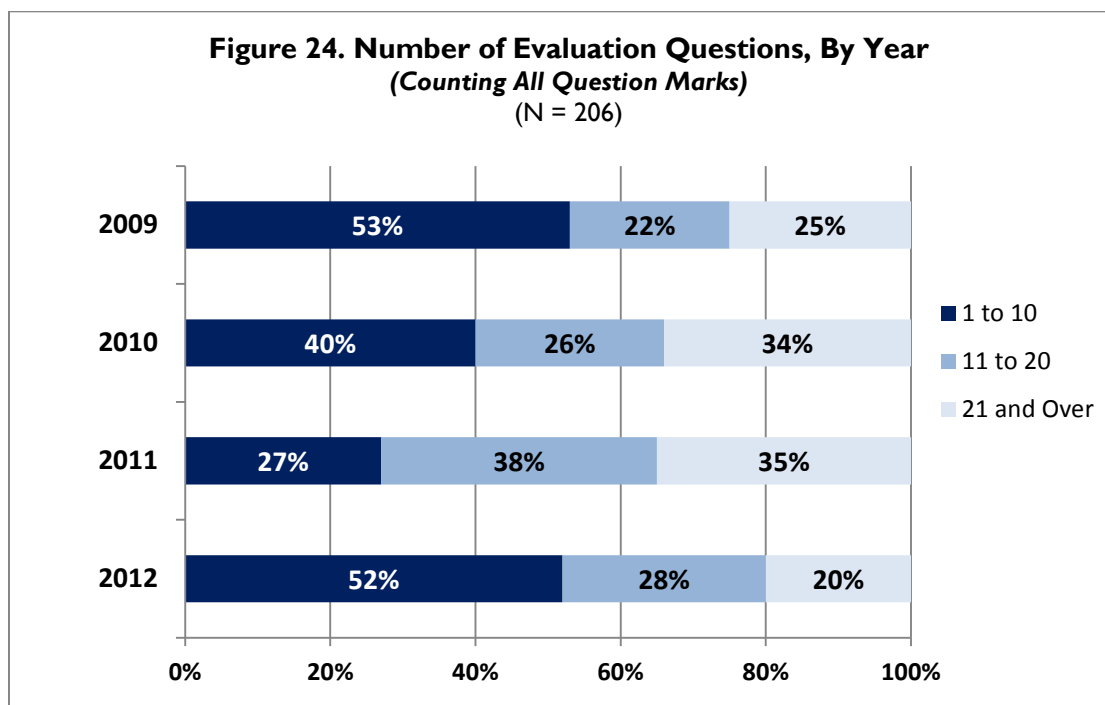
When considering which of the three methods of counting questions would be most appropriate for this meta-evaluation, MSI gave particular weight to the fact that Appendix I of the USAID Evaluation Policy states that “evaluation reports shall address all evaluation questions included in the scope of work” as an evaluation quality standard. MSI also recognized that every evaluation question has methodological implications for the collection of data and its analysis. This fact was also noted by a USAID regional representative who indicated, during a small group interview, that USAID staff do not always fully understand the methodological implications of the questions they are asking and how that impacts the cost of evaluations. With these considerations in mind, MSI elected to use the second option, the number of explicit question marks in a set of evaluation questions, as the basis on which to report numbers of evaluation questions for this study. Accordingly, figures in this section are based on a count of the question marks visible in main and sub-questions in evaluation reports and their associated SOWs.

Of the 215 evaluations reportedly having evaluation questions, written versions of those questions could only be located for 206 of those evaluations. MSI’s count of the number of 2009–12 evaluations that included actual questions is therefore based on that number.

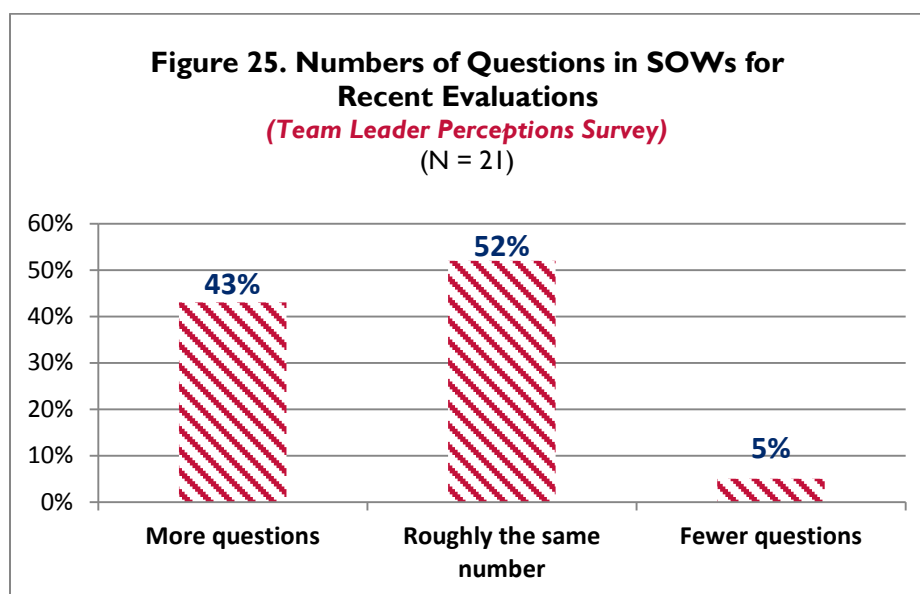
Figure 23 below shows that, when using a count of the number of explicit question marks in these 206 evaluations, 43 percent of USAID evaluations stayed within USAID’s “small number” guideline. Roughly equal percentages of evaluations fell into each of the other two clusters of numbers of questions described above.



On an annual basis, Figure 24 displays the percentage of evaluations falling into each of these clusters. As this figure indicates, a higher number of evaluations stayed within USAID’s “small number” guideline in 2012 than in the previous two years. This cannot be viewed as a recent improvement; however, since the percentage of evaluations meeting that guideline in 2009 was roughly the same as in 2012.



As the data presented in Figure 24 above suggests, there remains a great deal of variation in the number of evaluation questions included in evaluation SOWs. MSI's survey of recent evaluation team leaders provides further insight into this situation. Speaking from the experience of their most recent field evaluation work, team leaders who responded to the survey indicated that the number of evaluation questions appeared to be holding constant or rising rather than declining, as shown in Figure 25.



While annual data on the number of evaluation questions does not by itself demonstrate that USAID staff are trying to reduce the number of questions in evaluation SOWs, data from group interviews with USAID technical and geographic bureau staff do point to an emerging concern with the number of questions. In small group sessions, one regional representative claimed to have seen a decrease in the

number of questions and three other regional bureau representatives indicated that they now recognize that having fewer evaluation questions leads to better evaluations. This same view was articulated by a technical office representative in a separate small group interview. Raising a concern about guidelines on the numbers of evaluation questions, one representative of a firm that does evaluation work for USAID suggested that a forced decrease in the number of questions could result in more compound questions appearing in evaluation SOWs.

Number of Questions

“A small number of questions do not automatically lead to a good evaluation, but a long list always leads to a bad evaluation.”

USAID Regional
Bureau Representative

Figures 26 and 27 below are presented to facilitate an understanding of patterns with respect to numbers of evaluation questions by USAID regions and sectors.

Figure 26. Number of Questions, in Clusters, By Region
(Counting Question Marks) (N = 206)

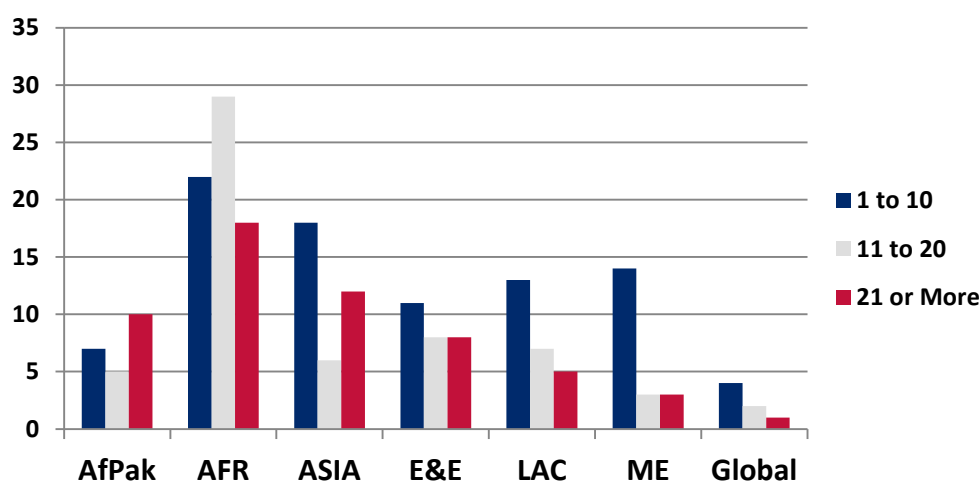
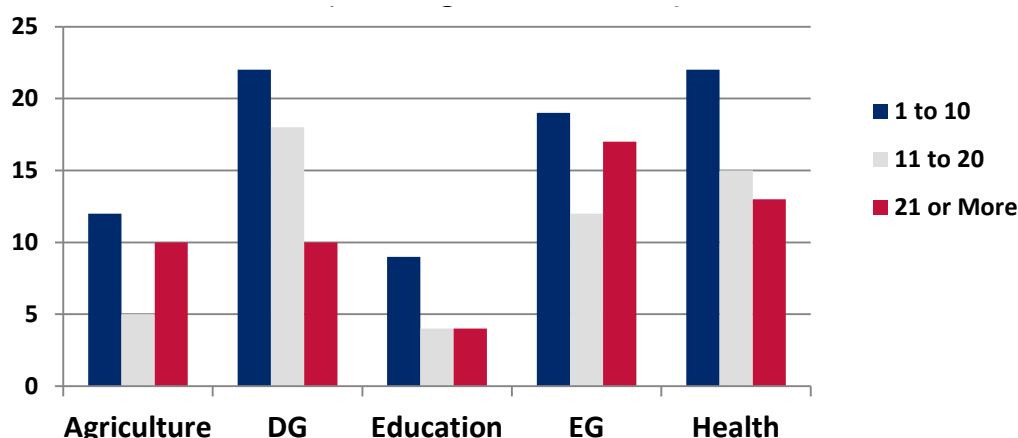


Figure 27. Number of Questions, in Clusters, By Sector
(Counting Question Marks) (N = 206)



With respect to the number of questions addressed, USAID Forward evaluations did no better than other evaluations in terms of addressing 10 or fewer questions. USAID Forward evaluations were instead two percentage points more likely to address more than 10 evaluation questions.

H. Team Composition

USAID staff identified the types of evaluation team members they encounter when preparing evaluation SOWs. USAID policy calls for special attention in SOWs to three types of evaluation team members. First, USAID's Evaluation Policy requires that team leaders for both performance and impact evaluations have external team leaders (i.e., individuals who do not work for USAID and have no relationship to the program or project they will evaluate). Second, since 2008 or earlier, USAID guidance has required that one member of every evaluation team be an evaluation specialist. Third, USAID evaluation guidance encourages Agency staff to include partner country professionals on evaluation teams. In this meta-evaluation MSI did not review evaluation SOWs to determine the frequency with which they complied with this guidance; rather, evaluations were rated on whether reports identified an evaluation team leader, an evaluation specialist, or the presence of local team members. It is important to note, in this regard, that evaluation reporting on team composition may understate compliance with USAID's guidance on team construction. A review of evaluation SOWs would be needed to assess this feature of evaluations directly.

I. Evaluation Cost

In the 1983 meta-evaluation of 270 USAID evaluations conducted by a team from Triton, evaluation cost information for 92 (34 percent) of the evaluations were located. They then compared the evaluation cost data with other study variables such as region, sector, and its score for evaluation completeness. Data on the cost of USAID evaluations in 1983 would have come from evaluation cover sheets used by USAID to transmit evaluations from the field to USAID/W; the last version of this form was called Form AID 1330–5 (10/87). A sample of the cost section from one of these forms, which USAID appears to have stopped collecting, is shown in Table 15.

Table 15. Sample Cost Section of USAID Form AID 1330-5 (10/87) Prepared in 1995

COSTS				
I. Evaluation Costs				
I. Evaluation Team		Contract Number OR TDY Person- Days	Contract Cost OR TDY Cost (US\$)	Source of Funds
Name	Affiliation			
1. Evaluation Team			\$114,321	NRM Project
John Clark (Team Leader)	University of Miami	45 person-days		
Peter Burbridge	Independent Consultant	42 person-days		
H. Soeparwadi	Bandung Inst. of Tech	36 person-days		
Krisnawati Suryanata	University of California, Berkeley	36 person-days		
2. Mission/Office Professional Staff			\$7,600	Operating Expenses (OE)
Dennis Cengel		21 person-days		
Ketut Djatl		21 person-days		
Agus Widiyanto		10 person-days		
3. Borrower/Grantee Professional Staff			\$3,350	NRM counterpart budget
Indah Dianti (BAPPENAS)		7 person-days		
Afrizal (MOFr)		5 person-days		
2. Mission/Office Professional Staff Person-Days (Estimate) <u>51 person-days</u>		3. Borrower/Grantee Professional Staff Person-Days (Estimate) <u>21 person-days</u>		

Reporting on evaluation costs in USAID meta-evaluations continued as a practice throughout the 1980s, making it possible for the 1987–88 meta-evaluations to provide USAID with comparative evaluation costs by regions, sectors, and types of evaluations, as illustrated by the cost by region table from that meta-evaluation below in Table 16.

Table 16. Cost of USAID Evaluations in 1987–1988

Cost of Evaluations by Bureau (1987–1988) in U.S. Dollars				
Bureau	Number of Reports	Mean	Minimum	Maximum
AFR	13	31,798	2,000	90,000
ANE	50	39,654	1,250	109,400
LAC	55	36,654	1,400	185,904
Other	9	40,900	8,601	107,568

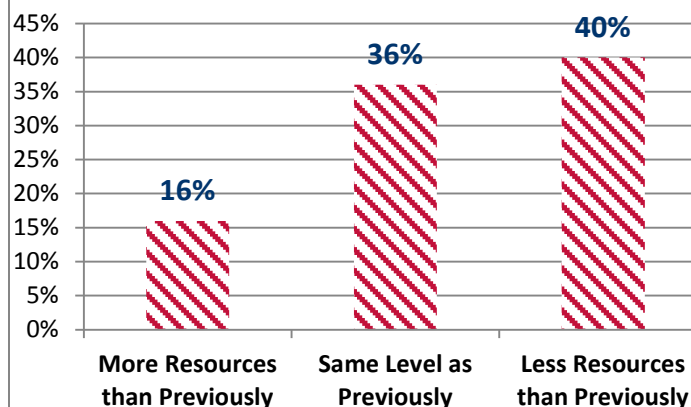
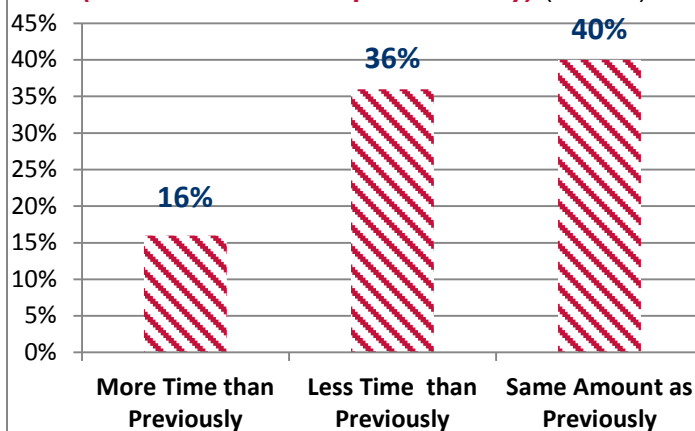
At some point, possibly around 1995 during USAID’s reengineering exercise that shifted its focus from projects to programs and eliminated the Agency’s requirement for a midterm and final evaluation of every project, USAID stopped using the evaluation transmission form for which the cost section is shown in Table 15 above. As a result, MSI and USAID faced challenges in trying to locate information on the cost and duration of evaluations conducted between 2009 and 2012. Responding to USAID’s offer to search its files, MSI identified 50 evaluations in the 2009–12 set that rated highest on the meta-evaluation review form, as well as 50 that rated lowest. For each of these evaluations, USAID searched contract files for evaluation duration and cost information.

After several unsuccessful attempts and much hard work, USAID was able to identify, using FactsInfo, information on a subset of the evaluations identified. Unfortunately, due to the misalignment of titles of evaluation reports with their associated contracts, inconsistent entry of data, and a shift in database structure part-way through MSI’s study period, it was determined that the data was unreliable and not usable.

While no quantitative cost data could be obtained for evaluations in the study period, MSI was able to obtain qualitative data through the Team Leader Perceptions Survey, which gauged the extent to which team leaders felt evaluation costs were changing. Among the 23 evaluation team leaders that responded to a question about evaluation resources, 40percent indicated that funds USAID made available for the most recent evaluations in which they were involved appeared to have declined compared with earlier evaluations they had conducted, as Figure 28 shows.

Evaluation Costs were also discussed in small group interviews with USAID staff and representatives of firms that conduct evaluations for USAID. With respect to the size of evaluation budgets, several views were presented. One regional office representative and one firm indicated that they had seen little change in the size of evaluation budgets, while four other firms and one regional office representative indicated that they have seen some larger budgets but that those were still insufficient to cover the work involved, as that too had increased. In a similar vein, one regional evaluation representative indicated that evaluation budgets are not reflective of the number of questions involved.

Another point made by a regional office representative was that to some degree the issue of evaluation budgets is being driven by Office of Acquisition and Assistance (OAA) which always chooses the offer with the lowest budget. In turn, that regional office representative said, firms are beginning to respond to that formula and make offers that may be priced below the budget levels that are actually needed.

Figure 29. Resources for Recent Evaluations
(Team Leaders Perceptions Survey) (N = 23)**Figure 28. Time for Recent Evaluations**
(Team Leaders Perceptions Survey) (N = 23)

J. Duration of the Evaluation

As with cost data, USAID was unable to locate any reliable quantitative data for evaluations in the sample, but MSI was able to obtain qualitative data on evaluation duration from small group interviews and from the team leader survey. In group interviews, one regional bureau representative and five firms indicated that insufficient time is being allocated for conducting evaluations. Two firms indicated that they would rather have fewer people on a team and more time available to them if they had the option.

Quality Suffers When the Time Allocated for Evaluations is Inadequate

“In the past five to seven years, and more broadly in the last decade, the amount of time to implement USAID evaluations has in general radically lessened given the tasks to be done. There seems to be an increasing urgency to get evaluations over and out of the way. Evaluations seem to be more and more carried out simply to get them over with; to satisfy less-than-adequate time frames for implementation; and to satisfy increasing monetary constraints. Despite good guidance from consulting firms, who also seem increasingly ‘squeezed’ by abrupt timing and planning changes over which they have little control, it seems increasingly evident to this consultant that financial constraints and fiscal control personnel (Contract Officers) with little knowledge of the evaluation issues have been controlling what is done. The quality of USAID evaluations suffers when consultants and firms are force-fitting too many activities and issues into increasingly inadequate time periods.”

Response to the Team Leader Perceptions Survey

The lack of adequate time to conduct evaluations was also noted by team leaders in the Team Leader Perceptions Survey, of which 36 percent of respondents said that less time was allocated for the most recent evaluation in which they were involved than had been the case in previous evaluations on which they worked. Figure 29, above, displays their responses on time allocated for evaluations. Team leaders also commented on the time allocated for evaluations in their narrative responses to the survey, indicating in several instances that they perceive a relationship between the time available for an evaluation and its quality.

2. EVALUATION QUALITY RATINGS IN DEPTH

Question I: To what degree have quality aspects of USAID’s evaluation reports, and underlying practices, changed over time?

Under this question, MSI examines how USAID evaluations have changed on a set of 37 evaluation quality factors and evaluation-related characteristics.* Evaluation quality factors included in the meta-evaluation instrument cover all key aspects of an evaluation report, including:

- Executive Summary—degree to which it accurately mirrors the most critical elements of the report
- Presentation of Project or Program Background—completeness from a reader’s perspective
- Description of the Project or Program’s “Theory of Change”—development hypotheses
- Evaluation Purpose—involving both a client’s focus on management needs and a team’s ability to capture and respond to management purposes
- Evaluation Questions—number, which is a client driven decision while team decisions affect whether questions are used to structure evaluation reports
- Team Composition—which clients describe in solicitations but which in practice are also affected by entities that form teams for clients
- Data Collection and Analysis Methods Used—specificity and links to questions
- Description of the Study Limitations
- Findings, Conclusions, and Recommendations—the adequacy of findings, clear distinctions between findings, conclusions, and recommendations and a logical, evidence-based flow from each one to the next
- Annexes—for their completeness

In this section, MSI reviews USAID evaluation compliance with evaluation quality expectations covered by the 37 quality factors in the meta-evaluation’s checklist instrument. Data for each quality factor are presented, and illustrate the percentage of evaluations that included the types of information and analyses recommended by USAID. These data are presented both as a whole, for the four-year period (2009–12), and on a year-by-year basis for each factor. In addition, at the end of each quality factor discussion, MSI summarizes performance information on each factor by region, sector, and whether or not there was a difference for that factor among USAID Forward evaluations and non-USAID Forward evaluations.

While many of the evaluation quality factors included in the meta-evaluation checklist, and reported on in this section, are determined by the efforts of an evaluation team, other factors are heavily influenced by the decisions of evaluation clients. Team composition and the number of evaluation questions to be addressed, for example, are factors heavily influenced by the decisions of evaluation clients.

To answer Question I, with respect to changes over time in the quality of USAID evaluation reports, MSI drew on several different sources described in the study’s methodology section, including:

- Evaluation quality ratings for 340 recent evaluations.
- Data on as many evaluation quality factors as possible from previous USAID meta-evaluations.

*MSI’s data collection instrument for this aspect of the meta-evaluation includes 39 items, but data for two of these items were deemed unreliable based on interrater reliability checks made during the data collection process.

- Responses from a small survey of 25 recent evaluation team leaders, which provide insight on a number of quality issues.
- Results of a series of group interviews with USAID staff and firms that undertake evaluation.
- Analysis of a subset of the meta-evaluation's data covering evaluations designated as USAID Forward evaluations, which may have changed at a different rate than other evaluations conducted in 2011 and 2012.

These data are used in an integrated manner to examine the frequency with which USAID evaluations in the sample complied with various evaluation quality prescriptions. Such prescriptions were provided in USAID's evaluation guidance materials, including the USAID Evaluation Policy introduced mid-way through the study period.

Before turning to a detailed review of quality factors included in the meta-evaluation checklist and the percentage of 2009–12 evaluations that rated positively on these factors, MSI presents Table 17, which summarizes the meta-evaluation's answer to USAID's question: To what degree have quality aspects of USAID's evaluation reports, and underlying practices, changed over time? This table shows, in Column 3, the net change in the average percentage of evaluations that complied with quality expectations on each of the 37 factors on the meta-evaluation quality factor checklist instrument. One additional quality factor, the percentage of evaluations that focused on (or were asked in their SOWs to address) 10 or fewer evaluation questions, was also included in the table as it was considered another factor in evaluation quality, though this data was collected through the basic evaluation characteristics instrument and not through the checklist.* The remaining columns on the right side of this table indicate whether the net change shown in Column 3, either positive or negative, was a straight line improvement or decline (linear), or whether there were fluctuations between 2009 and 2012. The final column, which indicated whether there was a steep rise in 2011–12 on any given quality factor, is intended to highlight factors where change may have been related to the issuance of USAID's Evaluation Policy in February 2011.

What this summary, presented in Table 17, shows is that:

- On 25 (66 percent) of the 37 evaluation quality factors MSI used to rate evaluations, there was a net improvement between 2009 and 2012. Two factors experienced no net changes, and ratings on 10 evaluation quality factors declined over the period, but only by one percentage point in half of those cases.
- Improvements were nonlinear for 83 percent of the 36 evaluation quality factors that changed over the four-year period covered by the study. Graphic representations of the quality improvement process resemble stock market charts on a busy day.

In short subsections below, MSI presents additional information on net changes over the meta-evaluation study period for each evaluation quality factor. Each of these short subsections examines both the percentage of evaluations that were rated positively over the four-year study period, and on a year-by-year basis. These factor specific subsections also indicate how the percentage of evaluations rated positively on a given factor may have changed by region or sector, or whether or not it was a USAID Forward evaluation in the period from July 2011 to December 2012.

Recognizing that regional, sectorial, and USAID Forward-related information is scattered over a relatively large number of pages devoted to quality factor analyses, MSI concludes its response to

*In the meta-evaluation, the number 10 was selected to represent the upper limit of USAID's ADS 203 recommendation, over most of the past decade, that evaluations address a "small number" of questions. In evaluation trainings conducted by MSI for USAID through 2010, and subsequently in USAID's EPM and EES courses provided in 2011–13, "up to 10" is the answer given by USAID participants in virtually all classes when asked to explain what USAID's recommendation meant by the term "small number."

evaluation Question 1 with a set of summary tables that will assist readers interested in any or all of these three foci. Readers with one or more of these special interests should feel free to review these tables at any time, as the overviews may help place information from the factor-by-factor discussions into context.

Table 17. Net Change in Evaluation Quality Factor Ratings Between 2009 and 2012

Evaluation Report Quality Factors (Full List)		2009–12 Evaluations - Net Percentage Point Increase/ Decrease	Pattern of Increase/Decrease 2009–12		
			Progression /Linear	Fluctuation /Nonlinear	Pronounced Increase in Last Two Years
#	Description				
More Than 10 Percentage Point Increase					
6	Questions in report same as in SOW	57		●	
33	SOW is included as a report annex	29			●
16	Study limitations were included	26		●	
35	Annex included data collection instruments	25		●	
12	External team leader	19		●	
30	Recommendations are specific about what is to be done	19		●	
18	Evaluation questions addressed in report (not annexes)	15		●	
22	Findings supported by data from range of methods	12			●
15	Report indicated Conflict of Interest forms were signed	12	●		
4	Management purpose described	11		●	
23	Findings distinct from conclusions/recommendations	11		●	
1 to 10 Percentage Point Increase					
39	Evaluation SOW includes Evaluation Policy Appendix I	8			●
29	Recommendations—not full of findings, repetition	6		●	
38	Report explains how data will transfer to USAID	5	●		
1	Executive summary mirrors report in all critical elements	5		●	
9	Data collection methods linked to questions	5			●
37	Statements of differences included as an annex	4		●	
17	Report structured to respond to questions (not issues)	4		●	
13	Report said team included an evaluation specialist	4		●	
7	Written approval for changes in questions obtained	4		●	

Evaluation Report Quality Factors (Full List)		2009–12 Evaluations - Net Percentage Point Increase/ Decrease	Pattern of Increase/Decrease 2009–12		
			Progression /Linear	Fluctuation /Nonlinear	Pronounced Increase in Last Two Years
8	Data collection methods described	3		●	
20	Social science methods (explicitly) were used	3		●	
31	Recommendations specify who should take action	2		●	
14	Evaluation team included local members	2		●	
2	Project characteristics described	1		●	
No Net Change					
10	Data analysis method described	0			
26	Alternative possible causes were addressed	0			
Net Percentage Point Decline					
32	Recommendations clearly supported by findings	- 1		●	
N/A	Number of evaluation questions was 10 or fewer	- 1		●	
5	Questions were linked to purpose	- 1		●	
34	Annex included list of sources	- 1		●	
27	Evaluation findings are sex disaggregated at all levels	- 1		●	
25	Unplanned/unanticipated results were addressed	- 1		●	
28	Report discusses differential access/benefits for men/women	- 2		●	
3	Project “theory of change” described	- 3		●	
24	Findings are precise (not simply “some,” “many,” or “most”)	- 7		●	
19	Reason provided if some questions were not addressed	- 11		●	
11	Data collection methods linked to questions	- 13		●	

A. Executive Summary Accurately Summarizes Critical Elements of Report Ratings

An executive summary is sometimes the only element of an evaluation report that USAID Mission Directors and senior managers in partner countries have the time to read. Where the wide distribution of the findings of an evaluation is desirable, but the entire report may not be necessary, an executive summary can be circulated as a stand-alone document. For these reasons, it is critical that an executive summary provide a “concise and accurate statement of the most critical elements of the report” (ADS 203.3.1.8).

Data from previous USAID meta-evaluations, in Table 18, show a steady increase in the proportion of evaluations that include an executive summary. For the period 2009–12, only 17 evaluations (5 percent) lacked an executive summary.

Table 18. Historical Presence of Executive Summaries

Presence of an Executive Summary	
Year	Percent
1983	49%
1985–86	68%
1987–88	64%
1989–90	59%
2009 only	90%
2009–12	95%

The current meta-evaluation went beyond simply looking at the presence of an executive summary, and further focused on whether executive summaries met USAID expectations with respect to including a “concise and accurate statement of the most critical aspects” of evaluation reports.

With regard to being concise, MSI found that among the 323 evaluations from 2009–12 that included an executive summary, the average length was nearly five pages. This length falls close to the outer limit set for future executive summaries in USAID’s 2012 How-To Note on Preparing Evaluation Reports.

For each executive summary, MSI also reviewed whether it accurately reflected what the evaluation found and whether it included information on specific evaluation report elements, provided they were also included in the evaluation report itself. Such elements include the evaluation purpose, background about the program/project being evaluated, the evaluation questions, a description of the study methods, a statement on study limitations (if any), the evaluation findings, and recommendations made by the evaluation team.* When rated on these quality criteria, of the 323 evaluations that included an executive summary, 147 (46 percent) were rated as having met expectations for their content. Generally speaking, when an executive summary did not receive a positive rating it was because one or more of the “critical elements” listed above was missing or because additional information was included in the executive summary that was not included in the evaluation report.

As Table 19 indicates, the percentage of executive summaries that received positive ratings was only slightly higher in 2012 than it had been in 2009. At the same time, Table 19 shows that in 2011, a much higher percentage (63 percent) of evaluations received positive ratings. The year-to-year progress shown in Table 19 is positive overall, but it is also nonlinear in nature. Average annual ratings on this factor fluctuated between 32 percent and 63 percent.†

*This set of critical elements differed slightly from the list on page 2 of USAID’s How-To note on evaluation reports in that it included recommendation and study limitations as critical elements.

†This was one of the largest fluctuations in average annual ratings between years and the only one not associated with a clear improvement in ratings between 2009 and 2012. While there was also a noticeable difference between ratings on this factor by sector, those differences did not directly parallel these year to year fluctuations.

Table 19. Executive Summary Accurately Reflects Report
(N = 323)

Executive Summary Mirrors Report		
2009 to 2012 Average Percentage	2009	42%
	2010	32%
	2011	63%
	2012	45%
46%		

Regionally, the percentage of evaluations rated as having an executive summary that reflected the most critical elements of evaluation reports paralleled the overall percentage for the study period—with two exceptions. A slightly higher average percentage of positive ratings for executive summaries were found for the ME region (52 percent) and for USAID/W (53 percent). Variations were somewhat wider on a sectorial basis with 60 percent of education evaluations receiving a positive rating for their executive summary quality, while evaluations of agricultural projects did less well on this dimension (38 percent). Other sectors fell somewhat closer to the overall average.

Ratings of the evaluation summary criteria used in the meta-evaluation show that 5 percent more USAID Forward evaluations were rated positively on this factor than non-USAID Forward evaluations.*

B. Project/Program Background Ratings

Since 2006, USAID evaluation guidelines on preparing evaluation reports have explicitly called for the inclusion of a description of the program or project, including the problem being addressed by the program or project.[†] When rating evaluations on this meta-evaluation element, MSI looked for the presence of information on aspects of program or project background highlighted by USAID, including the title of the program or project, the operating unit that managed it, its start and end dates, budget, implementing organization, geographic location, and target group. While not every element needed to be present, the coders were instructed that they must have a strong understanding of the program or project on which the evaluation focused.

Of the 340 evaluations rated, 90 percent included information on a range of background information factors listed above, as shown in Table 20. Ratings on this evaluation element fluctuated on an annual basis over a four percentage point range, improving slightly by 2012. Ratings also varied slightly by region, from 85 percent for E&E to 92 percent for AFR; and by sector, which ranged from 87 percent for DG to 96 percent for health. USAID Forward evaluations, with an average of 94 percent were higher than non-USAID Forward evaluations, which averaged 90 percent on this element.

*Table 69 at the end of this section summarizes differences between 69 USAID Forward evaluations and 85 non-USAID evaluations for the same time period on 37 meta-evaluation quality factors and the number of questions on which these evaluations focused.

†In addition to several editions of a USAID Evaluation Handbook that were issued as supplements to USAID Handbook 3 on programming between 1970 and 1990, and guidance in the ADS thereafter, USAID issued three guides to writing evaluation reports: *Constructing an Evaluation Report*, Blue and Hageboeck (2006), USAID *TIPS on Constructing an Evaluation Report* (2010), and the USAID How-To Note on Preparing Evaluation Reports (2012). Of the multiple editions of USAID's Evaluation Handbook issued over three decades, only the 1970 version is available on the DEC.

Table 20. Inclusion of Program/Project Background in Evaluation Reports
(N = 340)

Project Background Characteristics Described		
2009 to 2012 Average Percentage 90%	2009	90%
	2010	87%
	2011	90%
	2012	91%

C. Description of the Project or Program’s “Theory of Change” Ratings

Including the “theory of change” that underlies a program or project in an evaluation report helps readers understand how USAID addressed the problem or situation identified. This includes the intended outcomes of interventions and the hypotheses on which USAID based its expectation that the program or project will bring about those results. USAID’s 2011 How-To Note on Preparing Evaluation Reports expands on this description of “theory of change” by suggesting that if a Results Framework or Logical Framework exists, which documents USAID’s development hypotheses, that it be included in this section of an evaluation report.

Drawing on the “theory of change” description above, MSI found that 74 percent of evaluations between 2009 and 2012 had adequately presented the “theory of change” or development hypotheses underlying the program or project that was evaluated. As the annual data presented in Table 21 shows, average ratings on this meta-evaluation element fluctuated between 71 percent and 77 percent. Variation on a regional basis was somewhat broader, in which 80 percent of E&E evaluations and 69 percent of LAC evaluations included a description of the “theory of change” for the program or project evaluated. On a sector basis, inclusion of a “theory of change” was more likely for health projects (78 percent) than for EG projects (68 percent). Additionally, USAID Forward evaluations were four percentage points more likely to include a “theory of change” element than were non-USAID Forward evaluations.

Table 21. Inclusion of “Theory of Change” in Evaluation Reports
(N = 340)

Project “Theory of Change” Described		
2009 to 2012 Average Percentage 74%	2009	77%
	2010	71%
	2011	75%
	2012	74%

One other recent USAID study also examined how “theory of change” was handled in evaluations. This study, a review of foreign assistance evaluations completed in 2009 (Kumar and Eriksson), rated evaluations on whether the evaluation team had utilized a “theory of change” to structure its investigation. On this criterion, the study rated 26 percent of the 56 evaluations it examined as having utilized a “theory of change” to structure the evaluation research process or to present evaluation findings.

D. Evaluation Purpose Ratings

In evaluation reports it is generally expected that evaluation teams will restate what they understand to be the purpose or management reason for undertaking an evaluation. While a restatement of an evaluation’s purpose is not explicitly required by USAID, inclusions of this element in an evaluation

report has been encouraged by USAID guidelines on constructing evaluation reports since 2006. USAID's 2012 How-To Note on Preparing Evaluation Reports indicates that this section of a report should explain "why the evaluation is being conducted now, how the findings are expected to be used, what specific decisions will be informed by the evaluation, and who the main audiences are for the evaluation report."

Of the 340 evaluations reviewed for this study, 314 (92 percent) included a statement of the management purpose of the evaluation, while 8 percent of evaluations failed to include this element. Of the 314 evaluations that presented an evaluation purpose, 80 percent explained the management reason for undertaking the evaluation. Other evaluations that included a purpose statement were not scored as presenting a true management purpose; rather, they tended to simply say that the purpose was "to undertake an evaluation," or they described what was to be examined rather than explain why the project was being evaluated. These findings are consistent with findings from earlier meta-evaluations, in which 82 percent of USAID evaluation reports completed in 1989–1990 were also found to have stated a management purpose.

Table 22. Management Purpose Identified
(N = 314)

Management Purpose Described		
2009 to 2012 Average Percentage 80%	2009	70%
	2010	81%
	2011	86%
	2012	81%

As shown in Table 22, ratings on this element fluctuated between 70 percent and 86 percent on an annual basis, ending in 2012 with an 11 percentage point increase over the 2009 average rating. On a regional basis, ratings on the inclusion of a management purpose ranged from a high of 88 percent in the Middle East (ME) to a low of 75 percent in Africa (AFR). There were also differences by sector, in which 74 percent of health project evaluations included a management purpose compared with 91 percent of EG evaluations. In addition, MSI's review of USAID Forward evaluations found that this set of evaluations rated higher than non-USAID Forward evaluations by 13 percentage points on this evaluation element.

In addition to documenting specific management purposes cited in evaluations, MSI reviewed evaluations for references to USAID's 2011 Evaluation Policy statement about the primary purposes of evaluation in learning and accountability. It was found that these terms were rarely used explicitly, even in the most recent evaluations.

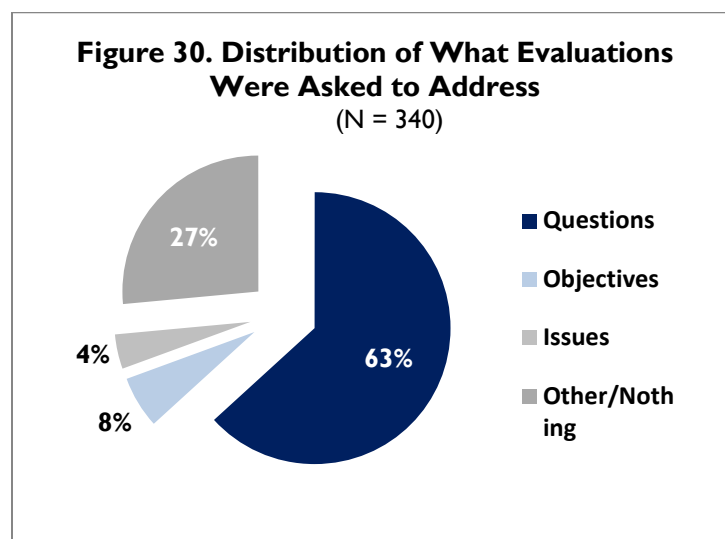
E. Evaluation Questions Ratings

Evaluation questions are most frequently the drivers of the evaluation process, not only for USAID but also for evaluations undertaken by other U.S. Government agencies and by most other donor organizations. Over the years, USAID has published evaluation handbooks, ADS, and guides for constructing evaluation reports, which have consistently indicated that evaluation questions should be stated in an evaluation report either as part of the evaluation purpose section or following the purpose in a separate section. These aspects of evaluation questions are addressed here.

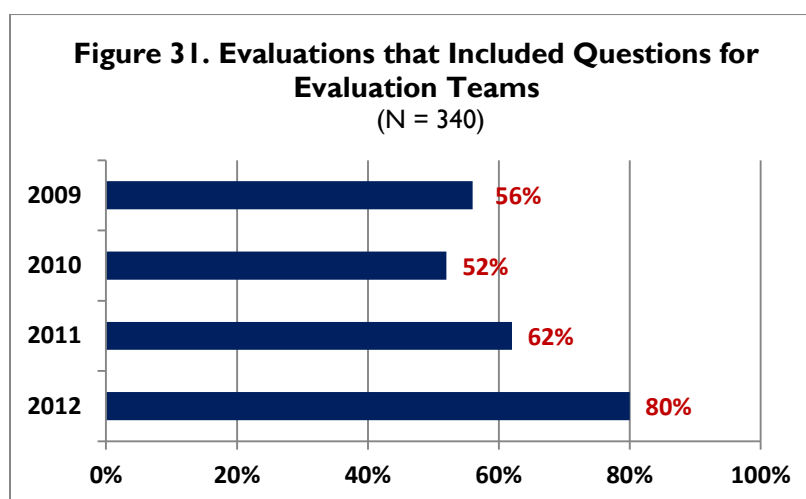
Presence of Evaluation Questions

The meta-evaluation's examination of 340 evaluations for the 2009–12 period found (as shown in Figure 30) that in 215 of these studies (63 percent), evaluation questions were either provided in the body of the report or their existence were alluded to in the report by referencing questions included in the evaluation SOW. For 27 of the remaining 125 evaluations, MSI found that items other than questions (e.g., issues or objectives) were identified for evaluation teams to address.

At the start of the study period, it was much more common to find that evaluations focused on issues or objectives than at the end of the study. This is possibly a function of USAID's guidance which states, in ADS 203.3.1.4, that evaluation SOWs should "identify a small number of key questions and specific issues answerable with empirical evidence." In contrast, USAID's Evaluation Policy and its evaluation training courses concentrate on questions as the starting point for an evaluation.



When data on the presence of evaluation questions were analyzed on an annual basis, MSI found that the presence of questions increased considerably during the last two years of the study period, from a low of 52 percent in 2010 to 80 percent in 2012, as shown in Figure 31. Although this shift was not explicitly called for in USAID's 2011 Evaluation Policy, that document does describe questions as an organizing framework for evaluations, as does USAID's 2013 How-To Note on Evaluation Statements of Work.



On a regional basis, the presence of evaluation questions also varied. Evaluations from Africa (56 percent) and USAID/W (54 percent) were less likely to include questions, even by reference, than were evaluations from AfPak (69 percent), Asia (69 percent), or ME (77 percent). Variations with respect to the presence of evaluation questions also existed on a sectorial basis, with evaluations in the health sector having the lowest incidence of questions being present (52 percent) and EG having the highest (74 percent). In addition, MSI found that USAID Forward evaluations were 26 percentage points higher than non-USAID Forward evaluations between July 2011 and December 2012 with respect to having evaluation questions for teams to address.

Questions Addressed in an Evaluation Report were the Same as in the Evaluation SOW

Prior to the release of the 2011 Evaluation Policy it was not uncommon to find evaluations presenting, and then addressing, a shortened list of evaluation questions taken from the full list of questions asked in an SOW. This practice sometimes involved bringing forward only the main questions and not the associated sub-questions, but other times it involved addressing only a portion of questions in the report when the full list of questions was particularly long. In some cases, it was explained that a shortened list of questions was the function of an agreement with USAID on where the evaluation would concentrate its efforts, while in other cases changes like this were made with no explanation. USAID's Evaluation Policy called for an end to such practices by requiring that all evaluation questions be addressed, which implicitly subsumes all sub-questions as well, though with proper documentation and approval from USAID the list of questions could still be shortened.

To determine how frequently evaluations have addressed questions included in SOWs in recent years, MSI included a rating factor on this issue. As shown in Table 23, of the 121 evaluations for which the meta-evaluation team had sufficient information to compare questions in the SOW with questions in the report, the meta-evaluation team found that 50 percent of the time questions in both places were identical. This percentage improved considerably from 2009 through 2012, but fluctuated in the years in between.

**Table 23 Evaluation Questions Addressed Were
Identical to the SOW**
(N = 121)

Evaluation Questions Addressed in Report were the Same as in SOW		
2009 to 2012 Average Percentage	2009	12%
	2010	50%
	2011	39%
	2012	69%

On a regional basis, evaluations from the AFR Bureau (59 percent) addressed the exact set of questions USAID listed in the evaluation SOW more frequently than did evaluations from other bureaus. On this factor, the low end of the range was represented by evaluations undertaken by USAID/W technical bureaus, which addressed the specific list of questions asked 25 percent of the time. USAID Forward evaluations (74 percent) were much more likely than non-USAID Forward (44 percent) evaluations completed during the last 18 months of the study period to have addressed specific questions asked.

In addition to requiring that evaluations address all the questions included in an evaluation SOW, USAID evaluation guidance requires that if the list of evaluation questions is changed in any way, permission in writing must be obtained from USAID. Data collected on this factor shows that of the 215 evaluations known to have been asked to address a list of questions, only four discussed receiving USAID permission to modify those lists questions.

Evaluation Questions Linked to Evaluation’s Management Purpose

USAID’s performance management system, which subsumes both performance monitoring and evaluation, envisions linkages between these two management tools, evaluation questions, and a management purpose to support evidence-based decision making. USAID’s ADS made the intended linkage explicit as early as 2003, describing evaluation as an “analytical effort undertaken to answer specific program management questions.”

This meta-evaluation included a question about the linkages between evaluation questions in studies carried out between 2009 and 2012 to assess how well this precept is integrated into USAID’s evaluation process and its evaluation reports. As Table 24 shows, in 99 percent of the 314 evaluations that included both a purpose statement and evaluation questions, a clear linkage between these two evaluation elements was found. This linkage was strong on an annual basis and on a regional and sectorial basis. The percentage of evaluations where this linkage was evidenced were high for both USAID Forward evaluations and for other evaluations completed in July 2011 and later.

Table 24. Evaluation Questions Were Linked to a Management Purpose
(N = 314)

Questions Were Linked to Evaluation Purpose		
2009 to 2012 Average Percentage	2009	100%
	2010	97%
	2011	100%
	2012	98%

In contrast to this quantitative finding on the strong linkage between evaluation questions and the management purpose of evaluations, one USAID regional representative stated in a group interview that evaluation questions are not always in line with the purpose of the evaluation.

The Quality of Evaluation Questions

Recognizing the difficulty of objectively rating the quality of evaluation questions, MSI’s meta-evaluation instruments did not include an evaluation question quality factor. Nevertheless, this topic did arise in small group discussions with USAID regional and geographic bureau staff and in discussions with representatives of firms that undertake evaluations for USAID. Comments offered by a few of these individuals on the quality of evaluation questions are summarized here.

In one group interview, three USAID technical office representatives expressed the view that the quality of evaluation questions has not improved over recent years. When this subject came up in the regional bureau small group meeting, one regional bureau representative expressed the same view while a second regional bureau representative disagreed, stating that the quality of questions had improved as a result of more serious reviews of evaluation SOWs and USAID staff putting more thought into their evaluation questions. This latter view was also expressed by one of the firms participating in a different small group interview.

F. Team Composition Ratings

Three aspects of evaluation team composition were examined through this meta-evaluation:

- Whether team leaders were external to and independent of USAID
- Whether an evaluation specialist was present on the team

- To what extent members of the evaluation team were local partner country nationals

Each of these team composition factors has a distinct history in USAID guidance, as summarized below.

Identification of Evaluation Team Leaders

Before turning to the discussion of external team leaders, it is important to bring to USAID’s attention an inadvertent finding concerning team leaders more generally. When working on the meta-evaluation’s Recent Team Leader Perceptions Survey, MSI found that the names of team leaders, whether external to USAID or not, could be found in evaluation reports for only 72 out of 184 (40 percent) of the 2011–12 evaluations. In other words, 60 percent of recent USAID evaluations failed to identify their study team leaders. This is an important quality finding which was not addressed by the quality factor checklist, but warrants reporting nonetheless.

External Team Leader

For decades, USAID evaluation guidelines have included a distinction between external evaluators and USAID personnel. Prior to the 2011 Evaluation Policy, USAID guidance did not require that evaluation team leaders be external to USAID, even though this was often the case. In the 2011 Evaluation Policy, USAID states that an “external evaluation is one that is commissioned by USAID, rather than by the implementing partner, and in which the team leader is an independent expert from outside of the Agency, who has no fiduciary relationship with the implementing partner.” Table 25 illustrates the frequency with which team leaders were identified in evaluation reports as being external to USAID. Between 2009 and 2012, there was a net increase of 19 percentage points on this evaluation element, despite fluctuations in the intervening years.

Table 25. External Team Leaders
(N = 340)

External Team Leader		
2009 to 2012 Average Percentage	2009	64%
	2010	78%
	2011	57%
	2012	83%
71%		

Regionally, the percentage of external team leaders, as best that MSI could determine, ranged from 55 percent in Asia to 84 percent in the AfPak region. On a sectorial basis, the involvement of an external team leader in evaluations was least prevalent in EG evaluations (56 percent) and highest among education sector evaluations (81 percent). USAID Forward evaluations had an average rating of 78 percent for the presence of an external team leader, which is 12 percentage points higher than non-USAID Forward evaluations. Data from two earlier studies help place current ratings of the external team leader’s element in context and is shown in Table 26 below. As early as 1983, over 50 percent of USAID evaluation team leaders were already external to the Agency, and that percentage continued to rise over time.

Table 26. Historical Data on External Team Leaders

External Team Leader	
Year	Percentage
1983	55%
1989–90	61%
2009–12	71%

Presence of an Evaluation Specialist on Evaluation Teams

Since 2008, USAID's ADS 203 has specifically called for the presence of an evaluation specialist on every evaluation team. The most recent update of this guidance in 2012 reaffirms this instruction. As was the case for identifying external team leaders, the meta-evaluation team's ability to accurately count the number of evaluation teams that had an evaluation specialist was limited because not all evaluations indicated if a team member was an evaluation specialist, or identified that individual by name if it was indicated. As Table 27 shows, data available in evaluation reports indicate that on average across the four-year meta-evaluation period, 14 percent of evaluations included an evaluation specialist on the evaluation team. On a net basis, this percentage rose between 2009 and 2012 by four percentage points. While quantitative data from the meta-evaluation did not reveal a particularly strong improvement on this evaluation element between 2009 and 2012, representatives of evaluation firms who participated in the meta-evaluation's small group discussions said that they are seeing many more requests for evaluation specialists and people with evaluation experience in recent solicitations, though one firm added that the description of evaluation specialists in evaluation SOWs is often very generic.

Table 27. Evaluation Specialists on Teams
(N = 340)

Report Said Team Included at Least One Evaluation Specialist		
2009 to 2012 Average Percentage	2009	15%
	2010	11%
	2011	8%
	2012	19%
14%		

Based on its review of past meta-evaluations, MSI notes that the highest percentage found for the presence of evaluation specialists on teams between 2009 and 2012 (19 percent) is still lower than a 27 percent figure cited in a meta-evaluation for 1998–99. However, authors of that study indicate that they inferred expertise, or its absence, from the methods used rather than give credit for evaluation expertise when an evaluation stated that at least one team member was an evaluation specialist.

On a sector basis, the frequency with which teams included an evaluation specialist was fairly similar, ranging from 11 percent for education and health project evaluations to 15 percent for agriculture and 16 percent for EG projects. The range was considerably wider on a regional basis as Table 28 shows. Also of note, USAID Forward evaluations were eight percentage points more likely to have had an evaluation specialist on the evaluation team.

Table 28. Reports That Identified an Evaluation Specialist on the Team, By Region
(N = 340)

USAID Regions and USAID/W						
AFPAK	AFRICA	ASIA	E&E	LAC	ME	USAID/W
20% of 35	16% of 128	13% of 55	7% of 41	5% of 42	12% of 26	31% of 13

Involvement of Locals on Evaluation Teams

USAID's 2011 Evaluation Policy encourages the participation of country partners on evaluation teams, stating: "To the extent possible, evaluation specialists with appropriate expertise from partner countries, but not involved in project implementation, will lead and/or be included in evaluation teams."

Historically, USAID evaluation guidance did not explicitly prioritize the use of local team members although around 2005 the Agency did begin to encourage Missions to involve local experts, local firms, and local nongovernmental organizations more widely in their approach to USAID projects and programs.

Based on evidence found in the 340 evaluations reviewed for the meta-evaluation, Table 29 indicates that 29 percent of the evaluations conducted between 2009 and 2012 included local team members. This percentage varied by year, but did not appear to increase significantly or in a linear manner between 2009 and 2012. MSI further found that its 2009 percentage for local team members on evaluations was comparable with the 33 percent local participation reported in USAID's meta-evaluation for 1998–99.

Table 29. Local Team Members Involved
(N = 340)

Evaluation Team Included Local Members		
2009 to 2012 Average Percentage 29%	2009	33%
	2010	25%
	2011	26%
	2012	35%

For the last 18 months of the meta-evaluation period, July 2011 to December 2012, there was a 10 percentage point difference between USAID Forward evaluations, of which 38 percent documented the involvement of local evaluation team members, and non-USAID Forward evaluations. On a sector basis, DG evaluations were more likely than other sectors to include local team members, with 34 percent of evaluations doing so. Evaluations in the EG cluster were the least likely to have included local team members (24 percent). On a regional basis, the range was somewhat wider with 22 percent of E&E evaluations indicating that local team members participated compared with 38 percent of ME evaluations as Table 30 shows.

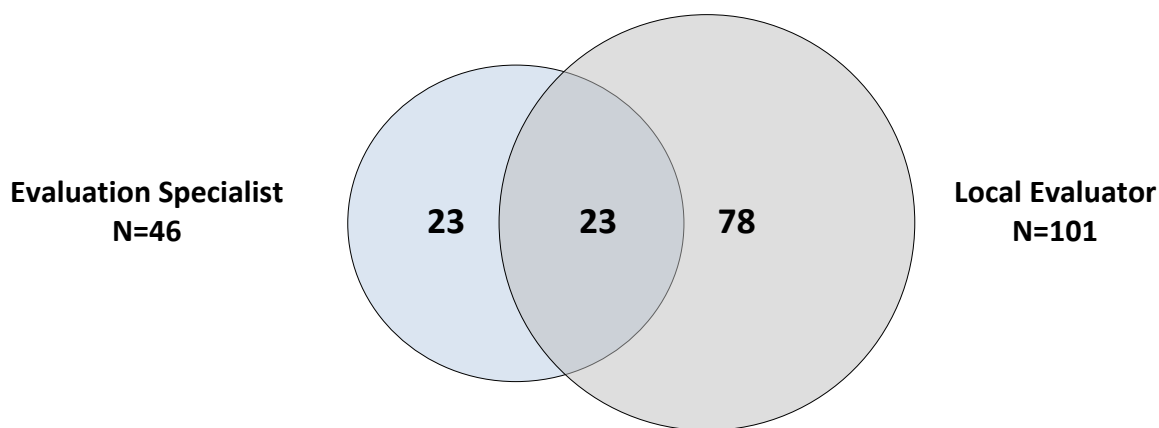
Table 30. Evaluation Reports That Identified Local Team Members, By Region
(N = 340)

USAID Regions and USAID/W						
AFPAK	AFRICA	ASIA	E&E	LAC	ME	USAID/W
37% of 35	29% of 128	35% of 55	22% of 41	29% of 42	38% of 26	0% of 13

Among the three aspects of team composition examined, the one identified in evaluation reports least frequently was the presence of an evaluation specialist, despite the fact that this factor has the longest history of prescriptive guidance in USAID ADS 203. After looking at the presence of an evaluation specialist separately from the involvement of local team members, MSI examined these two factors together. As the Venn diagram in Figure 32 below indicates, 124 of the 340 evaluation reviewed (36 percent) met one or both of these current criteria. This figure also highlights the fact that twice as many

evaluation reports identified local team members than indicated that the team included an evaluation specialist.

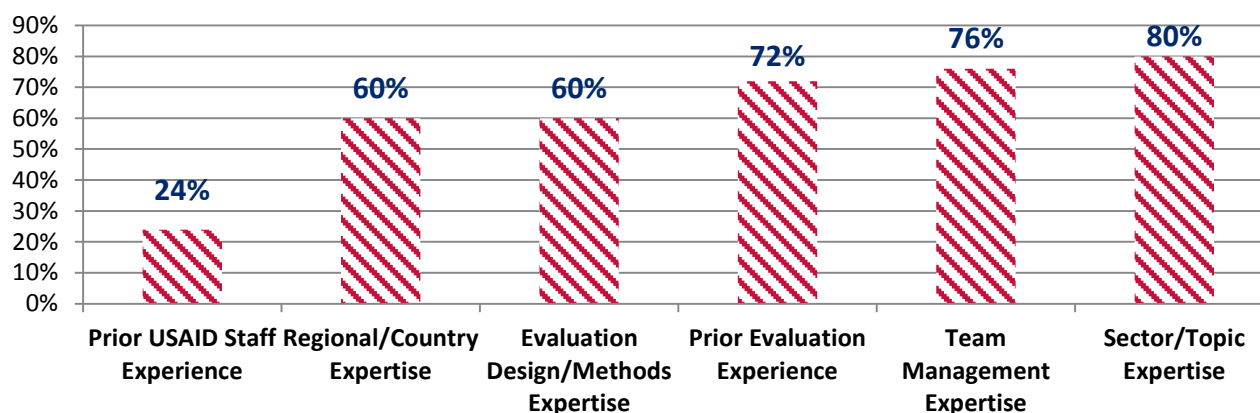
Figure 32. Presence of Evaluation Specialists and Local Evaluators on USAID Evaluation Teams



In small group interviews, participants commented on evaluation team composition requirements in SOWs. Two regional office representatives and one technical office representative noted that the recruitment and presence of technical specialists appears to take priority over evaluation expertise on teams, though the technical representative in these discussions also mentioned that in fields like health, the two are not mutually exclusive. In separate small group meetings, representatives of two firms that undertake evaluation work for USAID said essentially the same thing. During these discussions, one technical representative characterized this situation bluntly, saying that “the wrong people continue to write SOWs...they are sector specialists and not people who actually know evaluation.”

MSI heard in group discussions that USAID is most interested in having sector expertise when determining evaluation team composition. This theory was reinforced by data from a question in the Team Leaders Perception Survey that asked recent USAID evaluation team leaders to rate the factors they felt had the most influence on their being selected to lead evaluation teams. The distribution of their responses, shown in Figure 33, is consistent with USAID staff comments on the relative priority accorded to sector expertise in USAID’s evaluation team selection process.

Figure 33. Factors That Affect Selection of Evaluation Team Leaders
(Team Leader Perceptions Survey) (N = 25)



Commenting broadly on evaluation team composition in small group meetings, four firms that conduct evaluations for USAID and one technical office representative indicated that the quality of evaluation teams for recent evaluations has risen, including for evaluations being undertaken in 2013. In contrast, two USAID regional bureau staff said that they have not noticed any important changes in the composition of evaluation teams.

In these same small group discussions, two firms indicated the team composition requirements stated in SOWs are sometimes unrealistic, and said that in some cases no such candidate could possibly exist which meets all of the stated requirements. For their part, two USAID technical staff representatives said that the recruitment of personnel and the putting together of a roster is one of the largest challenges as available consultants are not always the best candidates for the job.

G. Team Awareness of USAID Evaluation Standards Ratings

USAID's 2011 Evaluation Policy recognizes that if USAID wants evaluations to meet higher quality standards, evaluators must be aware of those standards. To help raise awareness of USAID's evaluation standards among evaluation teams, the Evaluation Policy stated that evaluation SOWs "shall include" criteria for quality evaluation reports as found in Appendix I of the Evaluation Policy. This requirement was further reinforced when it was incorporated into ADS 203 in November 2012.

As the MSI team reviewed evaluation reports for 2011–12 in particular, it checked to see if this was being done. Among evaluation SOWs that were attached to 154 evaluation reports for these two years, the meta-evaluation found that 8 percent included either a copy of Appendix I or reproduced its work in SOWs, as Table 31 illustrates.

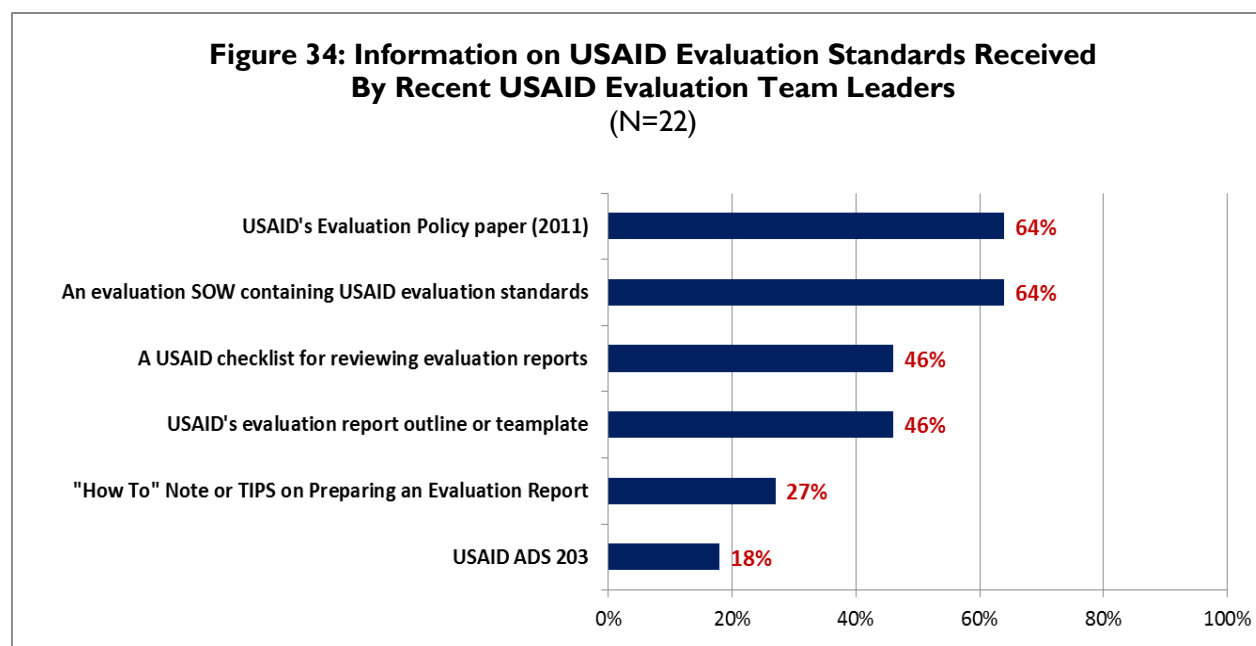
Table 31. 2011–12 Evaluation SOWs That Included Appendix I
(N = 115 Evaluations for 2011–12 with attached SOWs)

Evaluation SOW Includes Evaluation Policy Appendix I		
2009 to 2012 Average Percentage	2009	0%
	2010	0%
	2011	8%
	2012	8%
6%		

As a cross-check on whether evaluation team leaders for recent USAID evaluations were aware of USAID's evaluation quality standards at the time they conducted their most recent evaluation for USAID, MSI included a question on this subject in the Team Leaders Perceptions Survey. Data from this survey provides a more encouraging picture of USAID efforts to ensure that teams are aware of these standards:

- 22 of 25 (88 percent) of respondents to the evaluation Team Leader Perceptions Survey indicated that they had been provided with information about USAID's evaluation quality standards at the start of the evaluation period.
- Half of these respondents received this information directly from USAID, while the other half reported that the firm or NGO that had organized the evaluation team was the source of this information.
- Responses from 22 team leaders on this question indicate that most received more than one document relating to evaluation quality. USAID's Evaluation Policy was among the most frequently cited documents evaluation team leaders received.

Figure 34 displays the variety of documents that evaluation team leaders received concerning USAID evaluation standards prior to their most recent evaluation, some of which were being carried out in 2013.



H. Data Collection and Analysis Methods Ratings

Since 2008 or earlier, USAID ADS 203 has stated that evaluation SOWs, which are prepared by USAID staff, should identify the evaluation methods to be used in conducting an evaluation. USAID's 2012 update of this section of ADS 203 goes a step further and asks USAID to specify evaluation methods on a question-by-questions basis in an evaluation SOW:

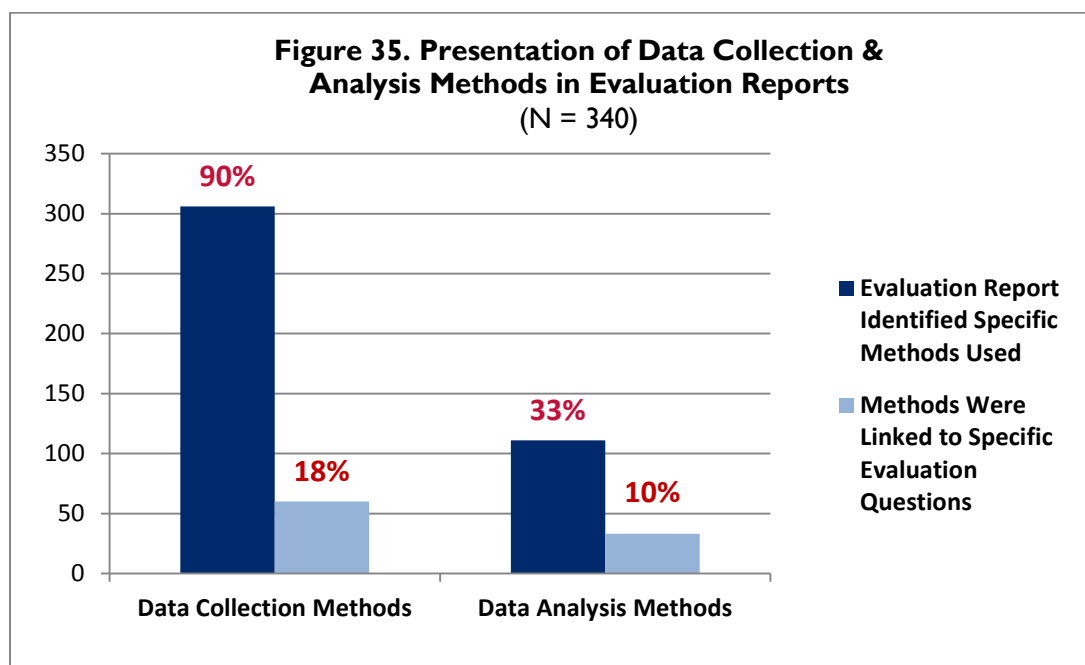
Identify evaluation method(s) that will generate the highest quality and most credible evidence on each evaluation question, taking time, budget, and other practical considerations into account and specify methods with sufficient detail.

For the 340 evaluations reviewed for the meta-evaluation, MSI rated each evaluation on whether it identified specific data collection and data analysis methods in the evaluation report or in a methodological annex. Evaluations were also rated on whether the evaluation methods described by teams were presented on a question-by-question basis, consistent with the intent of USAID ADS 203 quoted above.

As Figure 35 shows, a much larger proportion of evaluations specified their data collection methods (90 percent) than did those that specified their data analysis methods (33 percent). It may be worth noting that USAID's 2011 Evaluation Policy specifically calls for data analysis plans for evaluations, whereas previous guidance over the last few decades had not identified data analysis plans as a specific pre-evaluation requirement in the same manner that data collection methods had been required in advance of an evaluation. In this same vein, a representative of one of the firms that conducts evaluations for USAID, in a small group interview, raised data analysis methods as a topic and noted that USAID SOWs are increasingly including requests for data analysis.

On the question of whether USAID evaluations present data collection and analysis methods on a question-by-question basis (either in a matrix of the type promoted by USAID's evaluation courses or

by some other means), Figure 35 also shows that most evaluation reports from 2009 through 2012 did not do this for data collection methods (18 percent) or for data analysis methods (10 percent).



In the paragraphs below, choices made about data collection and data analysis in evaluations are examined in greater detail as are the frequency with which data collection and data analysis methods were described in evaluations on a regional and sector basis. Notably, there were few differences between USAID Forward evaluations and non-USAID Forward evaluations during the last 18 months of the meta-evaluation study period. Their performance with respect to describing data collection methods was virtually identical and on data analysis methods, non-USAID Forward evaluations received slightly higher rankings than USAID Forward evaluations.

Data Collection Methods

USAID evaluations tend to do well on this evaluation element. Overall, 306 (90 percent) of the 340 evaluations included descriptions of the data collection methods they used. There were, however, variations on an annual, regional, and sectorial basis. Annual data on the inclusion of a description of data collection methods in evaluations shows a low in 2010 of 80 percent, compared with a high of 95 percent in 2012. The net improvement over the four-year study period was three percentage points, as Table 32 indicates.

Table 32. Data Collection Methods Described in Evaluations
(N = 340)

Data Collection Methods Described		
2009 to 2012 Average Percentage 90%	2009	92%
	2010	80%
	2011	92%
	2012	95%

On a regional basis, ratings on this evaluation element ranged from 78 percent for evaluations carried out by E&E Bureau to 100 percent for evaluations undertaken by USAID/W. On a sector basis, 84

percent of DG and EG evaluations were rated as including descriptions of data collection methods. Higher ratings went to evaluations in education (96 percent) and health (95 percent).

As illustrated by Figure 35 above, the evaluations that described specific data collection methods generally did so without indicating which methods had been used to answer each of the evaluation questions the team addressed. USAID’s 2006 guide to Constructing an Evaluation discussed the importance of linking methods to questions and USAID 2013 How-To Note on Preparing Evaluation Reports reinforces USAID’s commitment to this practice. Since 2005 or earlier, a matrix approach to linking evaluation methods to evaluation questions has been included in USAID-funded evaluation courses for staff as an evaluation planning aid, and USAID’s 2011 volume on *Evaluation Statements of Work: Good Practice Examples* included a sample matrix, a version of which is shown below in Table 33.

Table 33. “Getting to Answers” Matrix

“Getting to Answers” Matrix					
Evaluation Questions	Type of Answer/Evidence Needed (Check one or more)		Methods for Data Collection, (e.g., Records, Structured Observation, Key Informant Interviews, Mini-Survey)		Sampling or Selection Approach (if one is needed)
			Data Source	Method	
1)	<input type="checkbox"/>	Yes/No			
	<input type="checkbox"/>	Description			
	<input type="checkbox"/>	Comparison			
	<input type="checkbox"/>	Explanation			
2)	<input type="checkbox"/>	Yes/No			
	<input type="checkbox"/>	Description			
	<input type="checkbox"/>	Comparison			
	<input type="checkbox"/>	Explanation			

In practice, the meta-evaluation found that only one of every four evaluations from 2009 through 2012 included a narrative description of the linkage between specific data collection methods and individual evaluation questions, or a matrix of the sort shown above. Not all evaluations identified specific evaluation questions, but among those that did, the presence of information on the relationship between data collection methods and specific evaluation questions in evaluation reports fluctuated over the meta-evaluation period, as Table 34 shows. Overall, there was an eight percentage point net gain on annual ratings on this evaluation factor by the end of the period.

Table 34. Evaluation Explained Linkages Between Data Collection Methods and Evaluation Questions
(N = 340)

Data Collection Methods Linked to Questions		
2009 to 2012 Average Percentage 18%	2009	11%
	2010	15%
	2011	24%
	2012	19%

For each of the 306 (90 percent) evaluations reviewed that actually described the evaluation’s data collection methods, the meta-evaluation extracted information on the specific methods those evaluations described. In addition, MSI raters examined the findings sections of reports and evaluation

annexes for evidence that the data collection methods discussed in their methods sections and annexes were actually used (i.e., data references made in the evaluation report emerged when those methods were utilized). What arose from this coding exercise was an inventory of data collection methods described and methods actually used. This inventory shows not only the frequency of use of certain methods, but also highlights the fact that some evaluations describe methods that they never actually use, and other evaluations use methods that they failed to describe in their methodology.

Table 35 below presents a profile of evaluation data collection methods found in USAID evaluations carried out from 2009 through 2012. Column 1 in this table lists data collection methods observed in evaluation reports. Column 2 indicates the number of evaluations that described these methods in their methods sections and annexes. Column 3 indicates the number of evaluations where the use of specific methods was verified, and Column 5 shows data collection methods on the basis of the percentage of evaluations in the study that actually used them. Between these two columns, Column 4 displays the difference between articulated plans to use data collection methods and actual use. As this column indicates, some methods were used much more frequently than evaluation methods sections in reports, taken alone, would have suggested.

Table 35. Planned and Actual Use of Evaluation Data Collection Methods

Collection Methods	Evaluation Described Plans to Use the Method	Report Review Found Evidence of Use of the Method	Difference Between Plan to Use and Actual Use	Percentage of Evaluations that Demonstrated Use of the Method
USAID Performance Data	243	285	+42 (117%)	84%
Document Review	252	274	+22 (109%)	81%
Key Informant Interviews	261	245	-16 (94%)	72%
Individual Interviews	187	185	-2 (99%)	54%
Unstructured Observation	156	152	-4 (97%)	45%
Survey	143	118	-25 (83%)	35%
Focus Group	147	100	-47 (68%)	29%
Structured Observation	24	26	+2 (108%)	8%
Group Interview	64	32	-32 (50%)	9%
Instruments (e.g., scale)	9	11	+2 (122%)	3%
Community Interview	5	3	-2 (60%)	1%

As the rank ordered list of data collection methods listed in Table 35 indicates, existing data, both project related and from secondary document reviews, are used in a large number of evaluations. This indicates that evaluators are building on what is learned from performance data, in line with aspirations outlined in ADS 203 for the relationship between these two management support activities. The next most frequently used set of evaluation methods includes several that are largely qualitative in nature, whereas more quantitative methods such as surveys (35 percent) are further down on the list of methods actually used.

MSI's finding from the methods analysis with respect to a heavy reliance on existing project and program performance monitoring data in USAID evaluations resonates with comments MSI received from individuals involved in the rating process—that what they were reading often seemed to have more of a final report character than what they would have expected from more comprehensive evaluations.

In small group discussions with USAID technical and regional bureau staff, as well as with firms that conduct evaluations for USAID, some participants offered their views on methods being used in USAID

evaluations. In a session with technical bureau representatives, four participants indicated that they could see virtually no change in data collection methods being used by evaluators. Meanwhile, one technical bureau representative along with four participants from an interview with firms indicated that the quality of methods used has increased. In addition, four representatives of firms indicated that USAID is now beginning to ask for more innovative and rigorous evaluation methodologies.

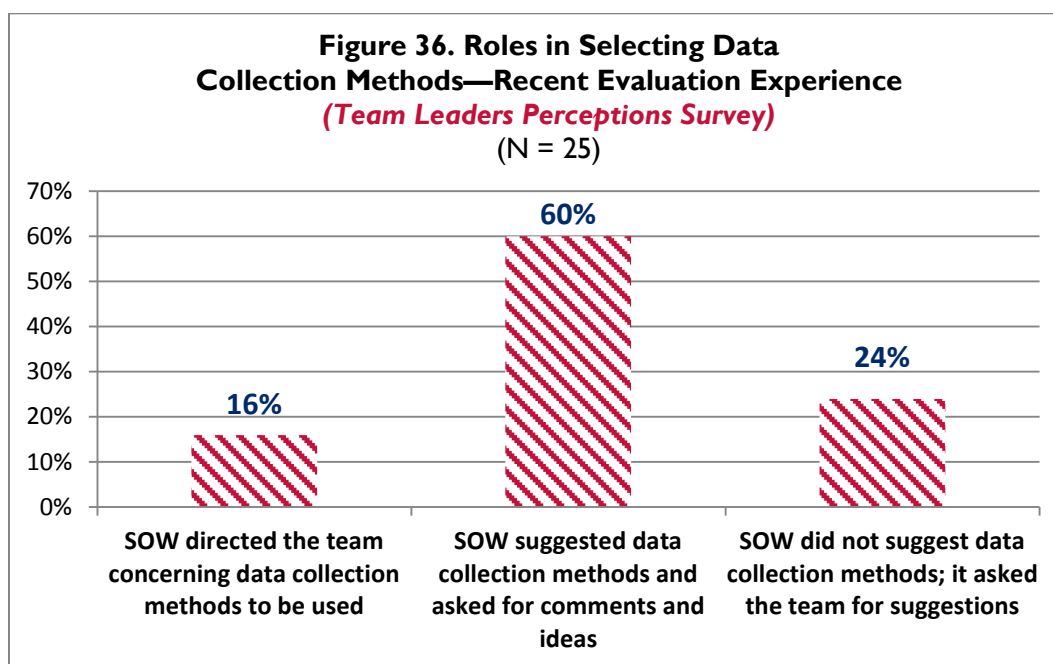
Consistent with some of these comments, MSI's review of previous USAID meta-evaluations indicate that most of the evaluation methods used in evaluations from 2009 through 2012 are the same as have long been used in evaluations. The mix of methods over time appears to have shifted towards more frequent inclusion of both surveys and focus groups, as Table 36 indicates.*

Table 36. Evaluation Methods Identified in Previous USAID Meta-Evaluations

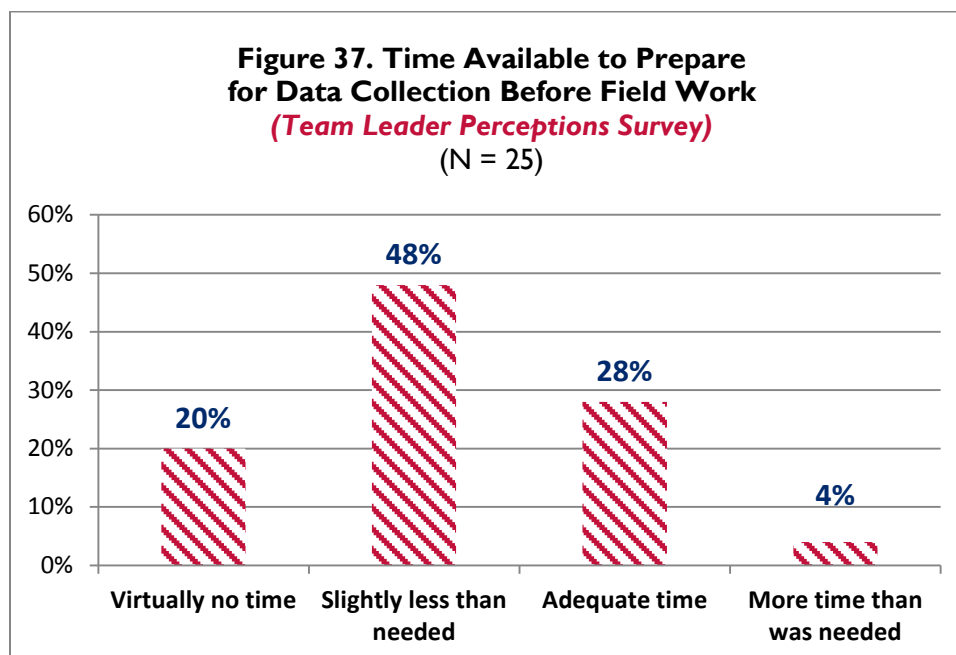
Data Collection Methods Used					
Year	Key Informants	Observation	Surveys	Focus Groups	Group Interviews
1997–98	89%	27%	10%	1%	5%
2005–08	100%	33%	39%	31%	53%
2009 only	80%	75%	51%	25%	21%
2009–12	72%	53%	35%	29%	9%

Additional insights on data collection methods used in USAID evaluations emerged from the Team Leaders Perception Survey MSI undertook as a complementary data collection method for the meta-evaluation. In this survey, team leaders for recent USAID evaluations were asked to characterize the coverage of evaluation methods in the SOWs they received for the most recent evaluations on which they worked. As Figure 36 indicates, most of the 25 team leaders who responded to this question described their SOWs as having suggested a set of methods. They also indicated that SOWs encouraged evaluation teams to respond to methodology suggestions and provide their own ideas about how to best go about answering evaluation questions.

*The term *focus group* was originally used to describe group data collection activities that could be characterized as involving a homogeneous population for reactive /opinion rather than fact gathering questions, and a facilitated process where the facilitator encouraged a discussion among participants, occasionally intervening to determine levels of consensus and shift from question to question. Over time, this term has come to be used more liberally to describe other types of group interviews with a specific focus. In this meta-evaluation, MSI accepted evaluation report statements to the effect that a focus group had been used, but did not attempt to classify what teams did as being “classic” focus groups or other types of group session, largely because the level of detail needed for this type of classification was not provided.



On the question of the time allotted to develop high-quality methods for evaluations, MSI queried team leaders about the adequacy of the time allotted in their agreements with USAID to prepare for data collection. This period is normally used to develop, translate, pretest, and modify instruments. Of the team leaders that responded to the survey, 68 percent said they had virtually no time or slightly less time than they needed for this evaluation task, as seen in Figure 37.



Data Analysis Methods

Only about one-third of USAID evaluations reviewed during the meta-analysis included a reasonably robust description of the methods evaluators used to analyze the data they collected using the methods described above. As Table 37 shows, the percentage of evaluations that did so was exactly the same at

the start and end of the study period, though it fluctuated in intervening years. MSI's review of USAID guidance over the years indicates that while the need to explain data collection methods in an evaluation report has been highlighted in a variety of guidance documents, less was said about data analysis. The identification of the need for "data analysis plans" for high-quality evaluations in USAID's 2011 Evaluation Policy is the most direct statement of this sort found.

Table 37. Data Analysis Methods Described in Evaluations, By Year
(N = 340)

Data Analysis Methods Described		
2009 to 2012 Average Percentage 34%	2009	34%
	2010	25%
	2011	37%
	2012	34%

As noted for descriptions of data collection above, there were also regional and sector differences with respect to the frequency of the types of data analysis evaluation teams used. Education evaluations (52 percent) most frequently included a description of data analysis on a sector basis, while EG evaluations (28 percent) did so the least. Regionally, evaluations in the E&E region (17 percent) included a description of data analysis least frequently and evaluations carried out by USAID/W (69 percent) did so most frequently.

Similar to data collection, only about one-third of evaluations that included a description of their data analysis procedures associated them with the specific evaluation questions they were used to address, as Table 38 illustrates. Notably, the percentage of evaluations that included a description of the relationship between data analysis methods described and specific evaluation questions addressed declined over the meta-evaluation period.

Table 38. Evaluation Explained Linkages Between Data Analysis Methods and Evaluation Questions
(N = 340)

Data Analysis Methods Linked to Questions		
2009 to 2012 Average Percentage 10%	2009	10%
	2010	10%
	2011	13%
	2012	6%

For data analysis methods, MSI conducted a coding and analysis effort similar to the one described above for data collection methods. Data analysis methods that evaluators stated they planned to use were extracted from evaluation methods sections and annexes and compared with information in reports to verify whether those methods had been used. Table 39 shows the results of this analysis. As this table shows, both descriptive statistics and content analyses of qualitative data appear to have been used more often than reports claimed they would be used. In the table below MSI highlights three types of analysis it tried to capture information on from evaluations.*

*In relation to this meta-evaluation question, descriptive statistics cover instances in which the MSI team found references to or examples of percentages, frequency distributions, cross-tabulations, or ratios. Inferential statistics subsumed procedures evaluators used to identify associations between variables, including correlation or regression analyses, as well as hypotheses testing statistics such as t-tests. Content analysis refers to qualitative data analysis procedures that involve pattern identification

Table 39. Planned and Actual Use of Evaluation Data Collection Methods

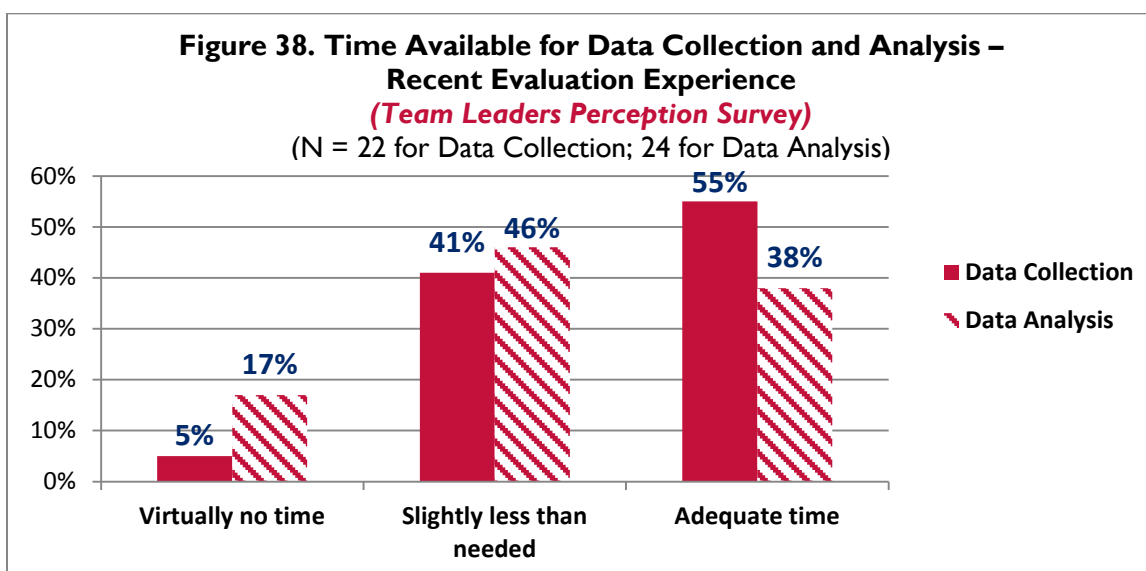
Data Analysis Methods	Evaluation Described Plans to Use the Method	Report Review Found Evidence of Use of the Method	Difference Between Plan to Use and Actual Use	Percentage of Evaluations that Demonstrated Use of the Method
Descriptive Statistics	91	215	+124 (236%)	63%
Content Analysis	86	93	+7 (108%)	27%
Inferential Statistics	29	29	None	9%

In addition to asking raters to indicate when evaluations used descriptive and inferential statistics, as well as content analysis of qualitative data, MSI used an open-ended rating question to identify any other types of data analysis that evaluations used. In particular, MSI was hoping to see examples of cost-effectiveness or unit cost analyses in evaluation reports, given the 2011 Evaluation Policy's specific encouragement to include methods that are suitable for collecting "financial data that permits computation of unit costs and analysis of cost structure" in programs and projects evaluated.

For 2009–12, MSI found no discussions of cost-effectiveness or unit cost analysis among the 340 evaluations it examined. There was one narrative discussion of cost-benefit analysis, though that appears to have been part of a discussion rather than a complete analysis. In contrast, MSI's review of prior meta-evaluation reports revealed that for the period 1987–89, 23 percent of the 287 evaluations rated undertook a detailed cost-effectiveness analysis. These data suggest that cost-effectiveness analysis may have been an important evaluation element earlier, which could be a function of guidance in the USAID Evaluation Handbook at that time and which was the predecessor to USAID's ADS 200 series. As early as 1970, USAID's Evaluation Handbook suggested that "progress indicators may be used to measure efficiency if they are used in such a way as to show the cost per unit in relation to the benefit accrued." This same 1970 USAID Evaluation Handbook included as a reference for evaluators, a 1967 book on cost-effectiveness analysis.

With respect to the adequacy of time allotted in evaluations for the actual collection of data and its analysis, the Team Leaders Perception Survey provides data regarding recent evaluations. As Figure 38 shows, evaluation team leaders generally feel that the time allotted for data collection is either adequate or slightly less than they need. The picture for data analysis differs, with a higher percentage indicating that virtually no time, or less time than is needed, is being allocated for data analysis.

including key words for phrases or other approaches to clustering similar responses obtained using questions and answer or observation techniques.



In addition to data culled from evaluation reports, the meta-evaluation also gained insights into data analysis in USAID evaluations from small group discussions, particularly with firms that conduct evaluations for USAID and from its survey of recent evaluation team leaders. In small group discussions, representatives of firms highlighted data analysis as an evaluation quality issue. One of these individuals remarked on an evaluation for which he said the “Mission gave zero days for data analysis.” Another commented, in a somewhat more positive vein, that “there is now more time for analysis, before there was not and it went straight from data collection to writing.” MSI’s team leader survey provided a broader view of current practices in this regard. As Figure 38 shows, recent team leaders continue to be concerned about the time available for data analysis, but some perceive this situation as improving. Evaluator descriptions of USAID evaluations—as providing too little time for data analysis—is consistent with the relatively low frequency with which MSI raters found the utilization of various data analysis techniques in 2009–12 evaluations.

I. Study Limitations Ratings

While the inclusion of study limitations is a longstanding “good practice” in research and evaluation, USAID has only recently begun to formally call for their inclusion by including such references in Appendix I of the Evaluation Policy and in ADS 203, subsequently. Nevertheless, MSI included a question about the presence of a statement of limitations in the checklist used to rate evaluations under this study. Overall the meta-evaluation found that over the four-year period of the study, an average of 51 percent of evaluations in the study sample included a statement of this type. What is more striking is the increase in the number of evaluations that did so between 2010 and 2011, as show in Table 40.

Table 40. Reports Includes a Description of Study Limitations

Percentage of Evaluation Reports that Include a Description of Study Limitations		
2009 to 2012 Average Percentage	2009	38%
	2010	34%
	2011	61%
	2012	63%
51%		

On a regional basis, USAID/W technical bureau evaluations (62 percent) included statements of study limitations more frequently than field Missions. Meanwhile, evaluations from Africa (59 percent) included statements of study limitations more frequently than did evaluations from AfPak countries (37 percent). On a sector basis, the spread was somewhat narrower, with 45 percent of health evaluations and 65 percent of agriculture project evaluations including study limitations. USAID Forward and non-USAID Forward evaluations received similar ratings on this factor.

J. Findings Ratings

Findings are the heart of an evaluation report and can affect the quality of an evaluation report in several ways. This section considers a number of factors that focus on evaluation findings and breaks them up into five groups, looking at:

- How findings were reported in relation to evaluation questions
- The relationship between findings and methods used
- Clarity of distinctions between findings, conclusions, and recommendations
- How gender was integrated into evaluation findings
- Whether findings covered broad evaluative concerns such as unplanned results and alternative possible causes of observed changes

Relationship Between Findings and Evaluation Questions

Over the past decade, USAID guidance has expressed a preference for the use of questions as the primary organizing framework for evaluations, and is consistent with evaluation “best practices” more broadly. Nevertheless, current and earlier editions of the *Planning Evaluations* subsection of ADS 203 (203.3.1.4.) have mentioned using both questions and specific issues as frameworks for conducting evaluations. Thus, not unexpectedly, Figure 23 earlier in this volume reported that 215 (63 percent) of the 340 evaluations focused on questions, while 27 percent addressed issues and 8 percent focused on evaluation objectives.

The choice between questions, issues, and objectives as the focus of an evaluation is made by USAID staff when a SOW is prepared. In this section, MSI examines how evaluation teams have responded to and structured their evaluations around evaluations when that is what they were asked to address. Subsections examine whether these evaluation reports presented their findings in relation to the evaluation questions, whether the questions they addressed were the same as questions in the evaluation SOW (rather than a shorter or different list), and whether these questions were addressed in the body of the evaluation report or relegated to an annex instead.

In each section below, we look only at the percentage of evaluations that scored positively on each factor. Ideally, every report would have been relevant for all factors, but this was not the case. For example, only 215 evaluations noted that the team had been asked to address a set of evaluation questions. Please note that the N, or number of relevant evaluation reports, for each factor may vary in the analysis, by factor. The N is provided in the table for each factor.

Presentation of Findings in Relation to Evaluation Questions

USAID guidance expects that findings described in evaluation reports will relate to the evaluation questions they were meant to address. This does not mean that every report should use a question-by-question structure. In some cases, it may be more appropriate to identify several questions for which findings are similar or related, and address that group as a cluster of questions and findings. In the meta-evaluation, raters checked to see if there is an approach or structure within the report that helped the reader to understand the relationship between findings and specific evaluations. If the relationship was

clear, the evaluation was rated positively, regardless of the exact structure of a report. This rating approach also helped distinguish evaluations that relate evaluation findings to questions from those that relate findings to issues, objectives, or some other framework.

Of the 117 evaluations MSI was able to rate on this issue, based on an ability to review both the questions asked and the way findings were presented, 54 percent were found to have linked findings to evaluation questions in the report structure or in some other transparent way. Annual data on this factor fluctuated and there was a modest improvement between 2009 and 2012 as shown in Table 41.

Table 41. Report Presented Findings in Relation to Evaluation Questions
(N = 117)

Reports was Structured to Present Findings in Relation to Questions		
2009 to 2012 Average Percentage 54%	2009	51%
	2010	48%
	2011	62%
	2012	55%

Of these same 117 evaluations, MSI found that evaluations from the LAC Bureau were at the low end of the range of evaluations on this factor while evaluations from the ME Bureau and USAID/W technical offices were both higher at 46 percent each. On a sector basis, health project evaluations were at the low end of the range in terms of findings that were clearly related to evaluation questions, while those from the E&E Bureau were higher at 43 percent. USAID Forward evaluations were also on the higher end of this range, with 48 percent presenting findings in relations to evaluation questions after July 2011, while non-USAID Forward evaluations for the same period were scored positively on this rating 40 percent of the time.

Location of Findings on Evaluation Questions in Report

The meta-evaluation checklist also included a factor that checked on where an evaluation question was addressed in an evaluation report (e.g., in the body of the report, annex, or elsewhere). Of the 232 evaluations for which the meta-evaluation had data on this factor, 62 percent addressed evaluation questions in the body of the report across the four years with an overall, but fluctuating increase, from 2009 to 2012 as Table 42 shows. On a regional basis, 68 percent of ME Bureau evaluations scored positively on this factor as did 50 percent of E&E Bureau evaluations. On a sector basis, 54 percent of DG evaluations were rated positively for addressing evaluation questions in the body of their reports, while 79 percent of EG projects also did this. USAID Forward evaluations did slightly better than non-USAID evaluation on this rating factor with 66 percent scoring positively, compared with 58 percent of non-USAID Forward evaluations for the same period.

**Table 42. Evaluation Questions were Addressed
in the Body of the Report**
(N = 232)

Evaluation Questions Answered in Body of Report, Not in an Annex		
2009 to 2012 Average Percentage 62%	2009	59%
	2010	71%
	2011	44%
	2012	74%

Findings in Relation to Methods

Three rating factors were used in the meta-evaluation to examine how a report presents findings in relation to the evaluation methods used.* The first of these focuses on whether evaluation reports based findings on the use of social science methods. The second looked at the degree to which reports used data from the full range of methods they described. The third factor in this cluster focuses on the precision with which findings were reported (i.e., number or percentages rather than general statements).

Findings are Based On the Use of Social Science Methods

Before turning to specific methods identified in evaluation reports, it is important to note that, in describing the types of methods to be used in future evaluations, USAID's Evaluation Policy calls for the use of social science methods. This term subsumes methods used in a number of disciplines and sectors including structured observations, survey research, key informant interviews, and the use of instruments such as weight scales or devices that measure length or distance, and other such tools. A broad rating factor in the meta-evaluation checklist was used to identify the percentage of evaluations that used these types of methods, irrespective of which particular method they used. On this factor, as shown in Table 43, MSI found that 77 percent of all evaluations appeared to use social science methods to obtain information. This percentage rose slightly between 2009 and 2012, but fluctuated downward in the intervening years.

Table 43. Findings Appear to Reflect the Use of Social Science Methods
(N = 319)

Findings Appeared to Reflect the Use of Social Science Methods		
2009 to 2012 Average Percentage 77%	2009	81%
	2010	64%
	2011	78%
	2012	84%

On a regional basis, evaluations from the ME Bureau (84 percent) had the highest average on this factor while those from the LAC Bureau (68 percent) had the lowest average. The spread was less wide on a sector basis with 82 percent of education sector evaluations and 73 percent of DG evaluations using these methods. There was only a slight, two percentage point difference in favor of USAID Forward evaluations on this factor during the last 18 months of the study period.

Findings Drew on Full Range of Methods Used

This rating factor was used to determine whether the findings sections in evaluation reports used data from all methods that the team stated they intended to use in the methodology section. For the 331 evaluations where it was possible to check findings against methods described, MSI found that 74 percent of reports drew on the full range of methods teams employed. Ratings on this factor improved by 12 percentage points over the study period in a straight line progression as Table 44 illustrates.

*MSI also attempted to collect data on a fourth rating factor in this grouping—whether the evaluation report provided a transparent connection between evaluation findings and the source of the data for those findings (e.g., 60 percent of the beneficiaries interviews reported that...; reanalysis of school records shows...; responses from mayors indicate that...). Unfortunately, in the process of checking interrater reliability MSI determined that the data for this factor was inconsistent and unreliable, and was therefore unusable for the purpose of this study. This factor was also removed from the checklists provided as annexes.

Table 44. Findings Clearly Drew on the Full Range of Data Collection Methods Used
(N = 331)

Findings Supported by Data from a Range of Methods		
2009 to 2012 Average Percentage 74%	2009	68%
	2010	71%
	2011	74%
	2012	80%

Among USAID regions, AfPak (82 percent) had the highest average rating for fully using the data from the range of methods reports described, while evaluations from the LAC Bureau had an average rating of 67 percent on this factor. On a sector basis, education evaluations (81 percent) were more likely to use data from the full range of methods described than DG evaluations (65 percent). USAID Forward evaluations after July 2011 had an average of 82 percent on this factor, while non-USAID Forward evaluations for the same period had a rating average of 71 percent.

Findings were Stated Precisely

While there is no specific USAID evaluation guidance that require quantitative findings be stated precisely rather than vaguely (e.g., “some,” “many,” or “most”), precision is considered to be good practice in evaluation and in most disciplines that utilize social science research methods. A rating factor on this issue was included to determine the degree to which evaluations reported quantitative data precisely. Of the 310 evaluations for which it was possible to apply this rating, 66 percent were found to have reported findings as numbers, percentages, or in other precise ways. Rather than improving over time, as seen in Table 45, annual ratings on this factor decline between 2009 and 2012.

Table 45. Quantitative Data Reported as Precise Numbers
(Not as “Some,” “Many,” or “Most”)
(N = 310)

Findings are Precise (Not Simply “Some,” “Many,” or “Most”)		
2009 to 2012 Average Percentage 66%	2009	74%
	2010	64%
	2011	63%
	2012	67%

Evaluations from USAID’s AfPak Region (81 percent) were found to precisely report findings more often than E&E Bureau evaluations (57 percent), which fall at the other end of the continuum. On a sector basis, the difference was also noticeable, but may be a function of the type of results produced. Among agriculture project evaluations, 78 percent reported findings precisely while DG projects, which tend to focus on qualitative results, were rated 48 percent on average in this factor. Notably, USAID Forward evaluations (71 percent) were a good deal more likely than non-USAID Forward evaluations (56 percent) to be precise in their reporting of evaluation findings.

Findings Are Distinguished from Conclusions and Recommendations

USAID evaluation guidance has long distinguished between findings, conclusions, and recommendations, indicating that each element represents an important step in a logical progression that moves from evidence to action. In evaluation reports, these elements are presented in various ways. Some reports

present findings, conclusions, and recommendations on a question-by-question basis while others cluster them around similar questions or use separate chapters in an evaluation report to cover each of these elements. Regardless of which structure an evaluation team chooses, it is important to indicate for the reader when shifts between these elements occur. By 2003, USAID ADS 203 had already spelled out what was to be included under each of these elements:

- Important findings (empirical facts collected by evaluators)
- Conclusions (evaluators' interpretations and judgments based on findings)
- Recommendations (proposed actions for management based on conclusions)

In rating evaluations on this factor, the meta-evaluation team did not look for any particular structure in which findings, conclusions, and recommendations were presented. Raters only needed to see that clear transitions were made between these elements in some way in order to rate evaluations positively. Using this criterion, MSI found that 41 percent of evaluations clearly distinguished between findings and recommendations, and between findings, recommendations, and conclusions when present. The percentage of evaluations rated positively on this feature rose in the period from 2009–2012, while fluctuating between these two years, as seen in Table 46.

Table 46. Evaluation Findings Were Distinguished From Conclusions and Recommendations
(N = 340)

Findings Were Distinct from Conclusions and Recommendations		
2009 to 2012 Average Percentage 41%	2009	37%
	2010	42%
	2011	37%
	2012	48%

Regionally, 54 percent of evaluations from the ME Bureau distinguished between findings and recommendations, while those from the LAC Bureau rated lower with 29 percent making these distinctions clearly. On a sector basis, agriculture project evaluations made this distinction (78 percent) more often than DG evaluations (48 percent). For the final 18 months of the study period, 71 percent of USAID Forward evaluations made these distinctions clearly when compared with 58 percent of non-USAID Forward evaluations.

Integration of Gender into Evaluation Findings

Since 2003 or earlier, USAID has required that data about people be disaggregated by sex for both performance monitoring and evaluations. In its 2012 update of ADS 203, USAID further clarified how gender is to be addressed in evaluations SOWs to which evaluation teams are expected to respond:

- Identify all evaluation questions for which gender-disaggregated data are expected
- Identify questions for which an examination of gender specific or gender differential effects are expected

This subsection reports on MSI's findings from the meta-evaluation on each of these gender dimensions.

Gender-Disaggregated Data

As indicated above, evaluations were rated as having adequately responded to USAID guidance on the sex disaggregation of evaluation data if they included sex disaggregated data in results, such as the adoption of new health, education, civic participation, or livelihood practices, as well as data on the participation of men and women in training programs about these practices. To only provide data on the

numbers of men and women trained was insufficient to garner a positive meta-evaluation rating on this factor.

Applying this requirement in the presentation of sex-disaggregated data on “people level” results (where sex disaggregation was both appropriate and potentially feasible), the evaluation found the percentage of evaluations that received positive ratings were roughly the same at the start and end of the meta-evaluation study period. The average for the four-year study period, as shown in Table 47, was 20 percent of 274 evaluations where people level results were presented that should have been disaggregated by sex per USAID guidance.

Table 47. Findings Were Disaggregated By Sex At All Levels
(N = 274)

Evaluation Findings Disaggregated by Sex At All Levels		
2009 to 2012 Average Percentage 20%	2009	23%
	2010	15%
	2011	23%
	2012	22%

Average ratings on evaluations varied considerably by region on this rating factor, with 38 percent of evaluations in the AfPak region scoring positively on sex disaggregation at all relevant results levels while 6 percent of E&E Bureau evaluations were rated positively on this factor. Ratings also differed by sector, with 40 percent of education evaluations rated positively on sex disaggregation of data at all relevant levels and 14 percent of EG evaluations receiving positive ratings. MSI further noted that sex disaggregation of evaluation data at all relevant results levels was not a strong feature of USAID Forward evaluations, of which 19 percent were rated positively on this factor (i.e., slightly below the overall average for the study period), while 25 percent of non-USAID evaluations for the same period (July 2011 to December 2012) were rated positively in the sex disaggregation of evaluation data.

MSI’s review of earlier meta-evaluations showed that evaluations were rated on sex disaggregation of evaluation data as early as the 1989–90 meta-evaluations. For those two years, 22 percent of evaluations were scored as including sex disaggregated data. This percentage may not, however, be comparable with percentages from the current meta-evaluation since the 1989–90 meta-evaluations may have used a lower standard (i.e., inclusion of sex-disaggregated data at any level of results but not necessarily at all relevant levels, which was standard for the 2009–12 meta-evaluations).

To understand the sex disaggregation of evaluation data over time in relation to USAID requirements, the meta-evaluation team reviewed what USAID guidance required at various points in time. As Table 48 shows, the ADS language on sex disaggregation of data has changed a number of times. In several revisions of the ADS (2003, 2008, and 2010), explicit references to evaluations were included in a MANDATORY ADS 203 section on reflecting gender considerations in performance indicators. USAID’s ADS 2012 language covers disaggregation for more than performance indicators, but it is not included in a section that is tagged MANDATORY, and the terminology shifted from sex-disaggregated data to gender-disaggregated data although these two terms do not have exactly the same meaning.

Table 48. ADS Guidance on Sex Disaggregation in Evaluations Over Time

ADS Date	ADS Section	ADS Guidance on Sex Disaggregation of Data that Explicitly References Evaluations (underlining added for purposes of this table)
2003	ADS 203.3.4.3	MANDATORY. Performance management systems <u>and evaluations at the SO and IR levels must include gender-sensitive indicators and sex-disaggregated data.</u>
2008	ADS 203.3.4.3	<p>MANDATORY. Performance management systems <u>and evaluations at the AO and project or activity levels must include gender-sensitive indicators and sex-disaggregated data when the technical analyses supporting the AO, project, or activity to be undertaken demonstrate that:</u></p> <ul style="list-style-type: none"> • The <u>activities or their anticipated results involve or affect women and men differently.</u> • If so, this difference would be an important factor in managing for sustainable program impact.
2010	ADS 203.3.4.3	<p>MANDATORY. In order to ensure that USAID assistance makes the optimal contribution to gender equality, performance management systems <u>and evaluations must include gender-sensitive indicators and sex-disaggregated data when the technical analyses supporting an AO, project, or activity demonstrates that:</u></p> <ul style="list-style-type: none"> <u>a. The different roles and status of women and men within the community, political sphere, workplace, and household (for example, roles in decision making and different access to and control over resources and services) affect the activities to be undertaken.</u> <u>b. The anticipated results of the work would affect women and men differently.</u> <p>Gender-sensitive indicators would include information collected from samples of beneficiaries using qualitative and quantitative methodologies <u>or an examination of the project impact on national, regional, or local policies, programs, and practices that affect men and women.</u></p>
2012	ADS 203.3.1.5	(6) <u>Identify all evaluation questions for which gender-disaggregated data are expected; also identify questions for which an examination of gender specific or gender differential effects are expected.</u>

Gender Specific or Gender Differential Effects of USAID Programs and Projects

As indicated in Table 48 above, the importance of capturing information on the differential effects of USAID projects and programs on men and women has been highlighted in the ADS since 2008 or earlier. In the meta-evaluation for 2009–12, MSI looked for this type of information. What it found was that 32 percent of evaluations “identified, discussed, or explained how men and women participated in or benefited from the program or project evaluated.” Upon closer review of this data point, MSI found that in most instances, the discussions were very cursory and often limited to an anecdote. Accordingly, data on this evaluation quality factor should be understood to have addressed gender effects in only the most minimal way and not with rich text or significant quantitative data as USAID’s Evaluation Policy and gender policy envision. That being said, MSI found that on an overall annual basis, performance on this factor fluctuated and there was very little change in the depth of information on gender effects presented over the course of the study, as Table 49 shows.

Table 49. Evaluation Questions Addressed Differential Access/Benefits by Gender
(N = 262)

Report Discusses Differential Access/Benefit for Men/Women		
2009 to 2012 Average Percentage	2009	42%
	2010	27%
	2011	23%
	2012	40%

On a regional basis, reporting on gender differential access and benefits was higher for AfPak (69 percent) than for the E&E Bureau (17 percent), possibly as a function of larger differences between men and women on some variables such as education, work outside the home, and other issues. On a sector basis, health projects and programs (23 percent) were at the lower end of the range of evaluations in terms of reporting on gender specific or differential access and benefits than were agriculture projects (48 percent). USAID Forward evaluations and other evaluations completed between July 2011 and December 2012 received similar ratings on this factor.

In addition to counting the frequency with which evaluations discussed gender specific or differential performance and outcomes, MSI raters extracted and saved these sections, and a simple content analysis was carried out to ascertain what types of methods and data were involved. What the content analysis showed was that sections on gender specific effects often included useful observations and insights, but generally speaking the data seemed to be limited (e.g., based on a single interview or one person’s comments in a focus group). Two examples provided below are illustrative of the way in which evaluations included in the 2009–12 meta-evaluation addressed differences between men’s and women’s relationship to USAID projects, and sometimes explained how obstacles to participation were overcome.

- In Afghan culture, men outside of a family are not allowed to enter a home if the man of the house is not present. To address such issues, the project employed seven Afghan women as meter readers in Jalalabad and Mazar. The project hired brother-sister teams, because they found it was not possible to hire women alone. These women initially began as meter readers, but now increasingly serve as customer care representatives, and are beginning to constitute a *de facto* female extension service as part of each social outreach program.
- It was noted that, in Uganda, where couples are targeted in Stepping Stones then the outcomes to gender based violence improve, some cases of stigma and domestic violence were reported

by women after returning home with project inputs or supplies which their male partners feared would expose their HIV status to neighbors and the wider community. (Focus Group discussions, Kalongo, Paimol). Stepping Stones' design, which consisted of gendered peer groups, was gender sensitive and in the context of decisions made during the group meetings; for example, the time for the meeting was decided after consulting women and men on the most appropriate time, giving the group members room to do their chores that would otherwise prevent them from participating if meetings were inappropriately timed.

Inclusion of Findings on Broad Evaluative Concerns

In addition to examining the way in which findings were presented in evaluation reports, the meta-evaluation team used two rating factors to assess whether and to what degree USAID evaluations in 2009–10 addressed issues that are widely considered to be evaluation concerns and that, among other things, differentiate an evaluation from performance monitoring. This section thus examines what the 2009–12 meta-evaluation found with respect to the treatment of a) unplanned or unanticipated effects of programs and project, whether positive or negative; and b) alternate possible causes of observed results of USAID-funded programs and projects in both performance and impact evaluations where questions about causality or attribution were addressed.

Unplanned Results of USAID Programs and Projects

In USAID's 2008 and 2010 versions of ADS 203, the unplanned results of USAID programs and projects were identified as something an evaluation might examine.* Information about unplanned results that managers might acquire through informal methods or from performance monitoring is also cited as something that might trigger an evaluation to help understand their implications. An example given by the ADS was how unanticipated results affected men and women, while another example might involve the unanticipated environmental consequences of agricultural practices. Including unplanned results in the scope of an evaluation has long been considered "good practice" in the evaluation field, as evaluations that do so tend to be more inclusive of all program or project results when they reach conclusions about the value or merit of a particular effort.

For 2009–12, as illustrated in Table 50, MSI found that 15 percent of the 340 evaluations examined discussed the unplanned results of the programs and projects they evaluated. Percentages varied on an annual basis, but overall there was not much change over the four-year period. Data for this period is lower than was reported in the 1989–90 meta-evaluation which found that 25 percent of the 268 evaluations rated in that meta-evaluation discussed unplanned positive or negative results.

Table 50. Evaluation Addressed Unplanned/Unanticipated Results
(N = 340)

Unplanned/Unanticipated Results were Addressed		
2009 to 2012 Average Percentage	2009	15%
	2010	11%
	2011	19%
	2012	14%

Differences were noted on both a regional and sector basis in terms of reporting on unplanned results in evaluations from 2009 through 2012. On a regional basis, 19 percent of AFR evaluations discussed unplanned results on the high end of this range, while 4 percent of ME evaluations did so on the low end

*USAID's 2012 update of ADS 203 no longer includes mention of unplanned results as a focus of evaluations.

of this range. Both DG and education projects represented the high end of the range, with 22 percent of evaluations in each of these clusters discussing unplanned results, while 10 percent of EG evaluation reports did so. For the last 18 months of the study period, more non-USAID Forward evaluations (18 percent) discussed unplanned results, than did USAID Forward evaluations (12 percent).

Alternative Possible Causes of Observed Results

As USAID's 2011 Evaluation Policy makes clear, the attribution of observed results to a USAID program or project above the output level requires sufficient evidence to support these kinds of claims. USAID introduced impact evaluations that use a counterfactual—an appropriately selected group or set of units that does not receive a specific USAID intervention—to determine what would have occurred in the absence of that intervention. Experimental and quasi-experimental designs are generally quite effective in isolating the effect of a specific intervention. Yet even when impact evaluation designs are used, it is possible that other factors in the program or project environment will have influenced outcome measures of interest as discussed in Part I of this volume.

Regardless of whether USAID carries out an impact evaluation that isolates improvements attributable to a specific USAID intervention or conducts a project performance evaluation that uses non-experimental methods to try to explain USAID's role in bringing about a change in an outcome indicator status it has detected, understanding other possible causes that may have contributed can help USAID managers plan forward. USAID's 2006 publication on *Constructing an Evaluation Report* stretched the list of reasons for examining alternative possible causes in evaluations well beyond cause-and-effect hypothesis testing questions, stating that alternative causes should be considered prior to reaching virtually all evaluation conclusions:

[It] is a critical part of the evaluation team's responsibility to explain, rather than just observe. For every finding, the team needs to discuss as many alternative explanations as possible. This means using various available forms of correlation tests, including cross-tabulations, regression analysis, factor analysis, and qualitative analysis, as appropriate, for every finding. These tools help the team test out as many plausible alternative explanations as possible before reaching closure on a particular finding.

Accordingly, the meta-evaluation team scored all evaluations on whether they discussed alternative possible causes of observed change. Overall, as Table 51 shows, 10 percent of the 340 evaluations rated for 2009–12 discussed alternative causes. This percentage varied little over the four-year period.

Table 51. Questions Addressed Alternative Possible Causes of Observed Results
(N = 340)

Alternative Possible Causes were Addressed		
2009 to 2012 Average Percentage	2009	10%
	2010	8%
	2011	11%
	2012	10%

On a regional basis, 23 percent of AFR evaluations included an examination of alternative possible causes compared with 4 percent in ME evaluations. At 15 percent, agriculture project evaluations were more likely to discuss other possible causes in evaluations than other sectors. There was no difference between USAID Forward and non-USAID Forward evaluations on this factor.

More important than region and sector, for this rating factor, is the extent to which evaluations that ask questions about causality include a discussion of alternative possible causes and the extent to which this factor was addressed in USAID impact evaluations versus performance evaluations. As reported earlier, 94 of the 340 evaluations rated addressed at least one question that asked about causality, of which six were impact evaluations and 88 were performance evaluations. Broadly speaking, the percentage of impact and performance evaluations that discussed alternative possible causes appears to have been very similar:

- Of the 11 impact evaluations examined, four discussed alternative possible causes (36 percent).
- Of the 329 performance evaluations examined, 88 included questions about causality and 29 (33 percent) of those discussed alternative possible causes.

Information that showed one-third of performance evaluations addressed questions about causality and examined alternative possible causes may be a positive finding. That these evaluations did so suggests that either such questions were part of their SOW, which was not always attached to their reports, or that evaluators exercised appropriate caution when drawing conclusions about causality unless reasonably strong evidence exists to support them.

K. Recommendations Ratings

MSI's review of evaluation recommendations examined four aspects of this element for each evaluation that was reviewed. These included whether recommendations were distinct from findings and conclusions, whether they were supported by findings, whether they were specific in nature, and whether they were clear about who should take action. As the tables and discussion below show, more than 50 percent of the evaluations reviewed were rated positively on three of these characteristics.

Recommendations Were Distinct From Other Aspects of the Report

For the first three years of the study period, evaluations received very similar ratings. In 2012, the percentage of evaluations in which recommendations stood alone and were not overburden by new or repetitive findings increased, rising above the four-year average of 59 percent to 64 percent as Table 52 shows.

Table 52. Recommendations Stood Alone in Report
(N = 319)

Recommendations—Not Full of Findings or Repetition		
2009 to 2012 Average Percentage 59%	2009	58%
	2010	56%
	2011	56%
	2012	64%

On a regional basis, the highest percentage of evaluations in which recommendations stood alone came from the AFR Bureau (65 percent), while the lowest percentage of evaluations appeared in the ME Bureau (38 percent). With a somewhat narrower spread, 68 percent of health sector evaluations had distinct recommendations unburdened by excessive findings, while 47 percent of EG evaluations had this feature. There was no difference between USAID Forward and non-USAID Forward evaluations on this factor.

Recommendations Were Clearly Supported by Findings

The inclusion of recommendations that are sufficiently supported by findings is one of the most important issues from a professional evaluation perspective. It not only assures readers follow a

transparent path from findings to conclusions to recommendations in an evaluation report, but also detects instances in which unsupported recommendations simply appear out of nowhere. In some cases, recommendations about findings that were not discussed earlier in an evaluation report are accompanied by a set of facts that is new to the reader. In these instances, the problem is poor structure or discipline in handling the findings → conclusion → recommendation chain of logic. In other cases, however, where recommendations appear without any apparent support from study findings, a degree of skepticism is warranted. Unsupported recommendations were marked down in the meta-evaluation ratings since, at best, it ask readers to act on faith rather than evidence and, at worst, could involve bias that readers are not able to easily detect, which is one of the reasons why it is considered unprofessional.

Among the 340 evaluation reports examined for this 2009–12 meta-evaluation, and as seen in Table 53, 80 percent scored positively on this factor, meaning that only recommendations that were supported by findings introduced in earlier parts of the report were included in the evaluation’s recommendations section. There were slight fluctuations over the study period, and the percentage in the last year (79 percent) was just below the four-year average.

Table 53. Recommendations Were Clearly Supported by Evaluation Findings
(N = 318)

Recommendations—Clearly Supported By Findings		
2009 to 2012 Average Percentage 80%	2009	80%
	2010	76%
	2011	83%
	2012	79%

On a regional basis, the meta-evaluation found that 88 percent of evaluations from the AfPak region scored positively on the inclusion of clearly supported recommendations. At the other end of the spectrum, evaluations from the LAC and ME bureaus both averaged 65 percent on this factor. The range was narrower on a sector basis with 71 percent of EG evaluations and 84 percent of DG evaluations scoring positively on this factor. USAID Forward evaluations had a slight lead of two percentage points over non-USAID Forward evaluations on this factor.

Recommendations were Specific About Actions to Be Taken

USAID’s Evaluation Policy calls for recommendations to be specific, practical, and action-oriented. The importance of clarity in evaluation recommendations is a concern throughout the development community as it affects utilization. In the mid-2000s, a review of evaluation utilization at the World Bank found that a relatively low rate of acceptance of evaluation recommendations by staff was tied to a lack of clarity and specificity in the way they were written. Subsequently, in its 2008 *Annual Review of Development Effectiveness*, the World Bank reported an improvement in the acceptance of evaluation recommendations that was attributed wholly to the efforts it made to ensure that evaluation recommendations are specific and actionable.

Among the attributes of recommendations USAID’s Evaluation Policy highlights as being desirable, specificity about actions to be taken was the easiest for the meta-evaluation team to code based solely on information provided in evaluation reports. This attribute was thus included in the meta-evaluation’s rating checklist. Evaluations from 2009 through 2012, on average, were rated positively on this factor 72 percent of the time as shown in Table 54 below. Annual averages fluctuate on this rating factor, ranging

from 58 percent in 2009 and to 77 percent in 2012. It should be noted that the trend between these years was not linear and the percentage of specific recommendations peaked in 2010.

Table 54. Recommendations Were Specific About Actions To Be Taken
(N = 318)

Recommendations—Are Specific About What Is To Be Done		
2009 to 2012 Average Percentage 72%	2009	58%
	2010	79%
	2011	72%
	2012	77%

Evaluations from the AFR and ME bureaus received an average rating for the period of 69 percent on this factor while Asia Bureau evaluations rated higher at 78 percent. On a sector basis, agriculture project evaluations were rated as including specific recommendations 68 percent of the time compared with education and EG evaluations which did so 74 percent of the time. USAID Forward evaluations (67 percent) were found to be less likely than non-USAID Forward evaluations (82 percent) to have included specific recommendations during the final 18 months of the study period in which they were compared.

Recommendations Were Clearly Directed to Specific Parties

In addition to requiring that evaluation recommendations be specific, practical, and actionable, USAID's Evaluation Policy and the 2012 ADS update both indicate that recommendations should also define responsibilities for taking action on individual or groups of recommendations. When rating 2009–12 evaluations on this factor, the meta-evaluation team found the 49 percent of evaluations did this as shown in Table 55. Ratings fluctuated on an annual basis and had a spike in 2011, but overall rose only slightly over the four-year period.

Table 55. Recommendations Were Clearly Directed to Specific Parties
(N = 318)

Recommendations—Specific as to Who Should Take Action		
2009 to 2012 Average Percentage 49%	2009	43%
	2010	45%
	2011	63%
	2012	45%

On a regional basis, evaluations carried out by the E&E Bureau (58 percent) and USAID/W technical bureaus (50 percent) were more likely than other bureaus, including the Asia Bureau (41 percent) to designate responsibility for evaluation recommendations. Non-USAID Forward evaluations (68 percent) did this more consistently than USAID Forward evaluations (82 percent) during the final 18 months of the study period.

L. Annexes Ratings

Six annexes or annex-related issues were included in factors the meta-evaluation rated, of which three were mentioned in USAID TIPS and other guidance over several years (e.g., include the evaluation SOW, a list of sources, and study instruments). Two other factors discussed in this section, for example, Conflict of Interest forms and an explanation of how study data sets will be transferred to USAID, are more recent requirements that came in with the 2011 Evaluation Policy. As both of these could be handled either as an annex or discussed in the body of the report, they were ranked positively wherever they appeared. The final element discussed in this section, the inclusion of a statement of differences, has long been a USAID evaluation report option, but this type of annex is optional and tends to be used only where serious differences exist.

Evaluation Statement of Work (SOW)

By 2003, USAID's ADS called for inclusion of the "scope" when documenting an evaluation and the 2006 guidance volume it published on *Constructing an Evaluation Report* explicitly stated that an evaluation SOW should be included as an annex to an evaluation report. For the four-year period covered by this meta-evaluation, 58 percent of evaluations included the evaluation SOW as an annex. That average improved each year over this period and was at 74 percent in 2012 as Table 56 shows.

Table 56. Evaluation SOW Was Included as a Report Annex
(N = 340)

SOW Included as a Report Annex		
2009 to 2012 Average Percentage 58%	2009	45%
	2010	38%
	2011	68%
	2012	74%

Data from meta-evaluations dating back to 1983 have recorded the frequency with which evaluation SOWs have been attached to evaluations and can be seen in Table 57. Interestingly, the highest percentage of SOWs being attached is 74 percent, which occurred in 1983, in 1989–91, and then once more in 2012, the final year of this study, even though the study average overall was quite a bit lower. Percentages for years between these three high periods were quite lower, reaching a low in 2010 of only 38.

Table 57. Historical Data on Evaluation Statements of Work (SOWs)

Presence of Evaluation Scopes of Work	
Year	Percent
1983	74%
1985–85	68%
1987–88	54%
1989–91	74%
1998–99	58%
2009–12	58%

On a regional basis, 48 percent of evaluations from the LAC Bureau included SOWs as annexes while 65 percent of those from the ME region did so. The picture was similar in terms of range on a sector basis

where 41 percent of education evaluations included the SOW at one end, and 65 percent of DG evaluations did so at the other end.

With regard to USAID Forward evaluations, 81 percent included the SOW as an annex compared with 69 percent of non-USAID Forward evaluations completed during the final 18 months of the meta-evaluation study period.

List of Evaluation Sources

Listing the sources of information used in an evaluation has long been considered a “best practice” even beyond USAID. A list of sources was recommended as an annex in the 2006 guide to *Constructing an Evaluation Report* published by USAID, in USAID’s 2010 TIPS with the same title, and its 2012 version of ADS 203. Normally, a list of sources includes documents used, and often presented in a bibliography, as well as a listing of groups or individuals with whom key informant interviews were conducted. Lists of sources do not typically include the names of survey respondents or focus group participants who contribute to the evaluation with the expectation of anonymity. Other interviewee’s names may not be disclosed if confidentiality agreements were signed or if evaluators recognize someone may become vulnerable should their names and views be too obviously connected to an evaluation report.

Among the 340 evaluations examined by the meta-evaluation team, 79 percent across the four-year period included a list of sources as described above. This percentage was slightly higher in all of the years of the study period except for 2010 where there was a noticeable decline in the percentage of evaluation reports that include sources as an annex (as shown in Table 58). On a regional basis, 63 percent of AfPak evaluations over the study period included a list of sources, as did 88 percent of E&E Bureau evaluations. On a sector basis the range was similar, with 67 percent of education and 82 percent of health evaluations including such lists. There was also a difference of about six percentage points between USAID Forward evaluations between July 2011 and December 2012 and non-USAID Forward evaluations for that same period. The USAID Forward evaluations did better on this rating factor with an average of 86 percent, which was also higher than the average for the final meta-evaluation year.

Table 58. Annex Included a List of Sources
(N = 340)

Annex Included a List of Sources		
2009 to 2012 Average Percentage 79%	2009	84%
	2010	68%
	2011	80%
	2012	83%

Evaluation Instruments Annex

Including an annex of evaluation instruments, much like including a list of sources, is considered an evaluation “best practice” well beyond USAID. Accordingly, it has been part of the guidance provided in USAID publications on *Constructing an Evaluation Report* in both 2006 and 2010. In 2012, instructions on including this type of annex were added to ADS 203. In principle, an instruments annex should include a copy of every instrument used. Thus, for example, if there were small surveys for four or five different subpopulations and the items differed by subpopulation, all five of these instruments would be included along with observation checklists, key informant interview schedules, focus group discussion plans, and other protocols for collecting data. The inclusion of all instruments is one of the things that make it possible for other research teams to replicate a study at a later point in time, or in another location. In

the meta-evaluation, MSI included two rating items. One focused on the presence or absence of an instruments annex. The second attempted to check on whether there was a one-to-one correspondence between methodologies discussed in the evaluation report and instruments used. The first of these ratings yielded reliable data, and is discussed below. The second proved difficult to execute as the methods sections did not specifically identify elements of their methods index or consistently connect them to descriptions of their data collection procedures.*

Among the 340 evaluations reviewed, MSI found that 61 percent of reports included an annex on study instruments. This percentage was higher in 2012 at 81 percent than in 2009 at 56 percent. This 25 percentage point increase was not, however, a stepwise progression, as Table 59 shows. On a regional basis, evaluations undertaken by technical bureaus in USAID/W included such annexes more often than other bureaus (77 percent). Evaluations from LAC represented the opposite end of this range, with 50 percent of evaluation reports including a methods annex with instruments over the four-year period. The range was somewhat narrower on a sector basis with 67 percent of health project evaluations and 55 percent of DG evaluations including an instrument annex. USAID Forward evaluations submitted during the last 18 months of the meta-evaluation period had a higher frequency (77 percent) than did non-USAID Forward evaluations for this same period (67 percent) with respect to the inclusion of this type of annex.

Table 59. Annex Included Evaluation Data Collection Instruments
(N = 340)

Annex Included Data Collection Instruments		
2009 to 2012 Average Percentage 61%	2009	56%
	2010	49%
	2011	55%
	2012	81%

Conflict of Interest Forms or Statement

While conflicts of interest are something USAID considers for all types of work it conducts, USAID's requirement that all evaluation team members sign Conflict of Interest forms or letters is new. The idea was introduced in the 2011 Evaluation Policy and was incorporated into the 2012 version of ADS 203. The meta-analysis team looked for the presence of Conflict of Interest forms or a statement about their availability, but did not expect to see many of these prior to 2011. In practice, this was precisely what the data show in Table 60. In 2011, one percent of evaluations included this information and by 2012 the percentage rose to 12 percent, suggesting an increased awareness of this new requirement. Given how new this requirement is, regional and sector data are not considered to be indicative of patterns of responsiveness on this rating factor. However, it is noteworthy that 13 percent of USAID Forward evaluations from July 2011 onward were rated positively on this factor, compared with 4 percent of other USAID evaluations completed during that same period.

*Though data was collected on this quality element, MSI determined that the data were inconsistent and unreliable, therefore this study will provide no analysis for this factor and the element has been removed from the checklists provided as annexes at the end of the report.

Table 60. Report Included Conflict of Interest Forms or Indicated That They Were Available
(N = 340)

Report Indicated Conflict of Interest Forms were Signed		
2009 to 2012 Average Percentage 4%	2009	0%
	2010	0%
	2011	1%
	2012	12%

Also, with regard to USAID's new requirement for Conflict of Interest statements or signed forms, MSI's survey of 25 team leaders for recent USAID evaluations revealed that 61 percent of them had been asked to sign a Conflict of Interest form before embarking on their most recent evaluation. This percentage suggests that the overall percentage of evaluations that include such forms may increase in future years.

Transfer of Evaluation Data Sets to USAID

Like the conflict of interest rating item, USAID's requirement for the delivery of evaluation data sets was also introduced in the 2011 Evaluation Policy. The 2012 update of ADS 203 makes this requirement more specific, stating that what is to be delivered to USAID includes "raw quantitative data and any code books." The meta-evaluation included an item on this factor, largely to detect whether this change in guidance produced a response. As Table 61 shows, the frequency with which evaluation reports indicated that data sets have been delivered to USAID has begun to rise. On a regional basis, AfPak evaluations rated highest on this factor with 6 percent of evaluations documenting the delivery of data sets. On a sector basis, 7 percent of DG evaluation reports discussed the delivery of data sets, and is the high end of the range from this perspective. USAID Forward evaluations (6 percent) were only slightly more likely than non-USAID Forward evaluations (4 percent) to have discussed data transfer.

Table 61. Report Indicated How Data Obtained by the Evaluation Will Be Transferred to USAID
(N = 340)

Report Explains How Data Will Transfer to USAID		
2009 to 2012 Average Percentage 2%	2009	0%
	2010	1%
	2011	2%
	2012	5%

Statement of Differences

The inclusion of a statement of differences in evaluation reports differs from other rating factors as this is an option rather than a requirement. The inclusion of a statement of differences prepared by team members or stakeholders (e.g., an implementing partner) has long been an option for USAID evaluation reports. Such statements can be found in evaluations conducted as far back as the 1980s, including in a series of evaluations carried out during that period by USAID's then Office of Evaluation. USAID published guidance on *Constructing an Evaluation Report* in 2006, the subsequent 2010 TIPs by this same name, and the 2012 update of ADS 203, indicating that statements of differences can be included as annexes. Table 62 shows an increase in the frequency with which reports included this feature toward

the end of the meta-evaluation period. Numerically, there were only 15 such annexes across 340 evaluations, but half of these were included in evaluations completed in the final year of the meta-evaluation study period.

**Table 62. Evaluation Included a
Statement of Differences as an Annex**
(N = 340)

Statements of Differences Included as an Annex		
2009 to 2012 Average Percentage 4%	2009	3%
	2010	2%
	2011	4%
	2012	7%

M. Summary Tables on Quality Factors

This section includes four summary tables prepared for readers who have a particular interest in seeing how evaluations performed on the 37 factors included in the meta-evaluation quality factors checklist, plus an unnumbered factor on whether evaluations addressed 10 or fewer evaluation questions. In turn, these tables summarize findings on performance by region and sector and for USAID Forward evaluations.

Percentage of Evaluations by Region Rated Positively on Evaluation Quality Factors

Table 63 provides a quick review of the percentage of evaluations in each USAID region, as well as those carried out by USAID/W technical bureaus, that were rated positively on evaluation quality factors examined in detail above in response to the question: To what degree have quality aspects of USAID's evaluation reports, and underlying practices, changed over time? Table 63 below introduces the quality rating factors in the order in which they are found in the meta-evaluation checklist (Annex C) and in the coders handbook, which explains how each factor was coded (Annex C).

In table 63 below, two pieces of information are included that were already presented in the factor-by-factor analyses in the preceding pages. These pieces of information are the number of evaluations for which the meta-evaluation had information on each factor, and the average percentage across regions that rated positively on each factor within the study time period; these data can be found in Columns 3 and 4 respectively. The remaining columns on the right side of the table present the percentage of evaluations over this four-year period that were rated positively in each region: Afghanistan and Pakistan (AfPak), Africa (AFR), Asia, Europe and Eurasia (E&E), Latin America and the Caribbean (LAC), the Middle East (ME), and USAID/W technical bureaus. A useful way to review the status of any given region on this table is to compare the percentage of its evaluations that was scored positively with both the overall average (provided in Column 4) and to the percentage for the region that has the highest compliance rating for that particular quality factor.

Under each region name at the top of Columns 8–11, the total number of evaluations from that region is also displayed, but it is important to note that percentages shown in this table are not necessarily percentages of the total number of evaluations from a region. Wherever the number of observations shown in Column 3 in this table is less than 340, the number of evaluations from any particular region will also likely be less than the total shown at the top of the region column.

Table 63. Quality Factor Ratings by Region*
(among those evaluations on which data on a factor were available)

Rating Factor		N =	Average All Regions	USAID Regions and USAID/W						
#	Description			AfPak (35)	AFR (128)	ASIA (55)	E&E (41)	LAC (42)	ME (26)	USAID/W (13)
1	Executive summary mirrors report all critical elements	323	45%	41%	45%	44%	44%	54%	47%	52%
2	Project characteristics described	340	90%	89%	92%	91%	85%	92%	86%	88%
3	Project “theory of change” described	340	74%	71%	75%	73%	80%	77%	69%	77%
4	Management purpose described	314	80%	87%	75%	87%	81%	75%	76%	88%
5	Questions were linked to purpose	190	99%	100%	98%	97%	100%	100%	100%	100%
6	Questions in report same as in SOW	121	50%	40%	58%	50%	42%	25%	50%	53%
7	Written approval for changes in questions obtained	62	6%	11%	5%	25%	0%	0%	0%	0%
8	Data collection methods described	340	90%	91%	92%	89%	78%	100%	90%	92%
9	Data collection methods linked to questions	265	23%	11%	25%	22%	17%	36%	26%	21%
10	Data analysis method described	340	33%	34%	38%	27%	17%	69%	29%	27%
11	Data analysis methods linked to questions	105	31%	9%	30%	43%	29%	33%	42%	33%
12	External team leader	281	71%	84%	73%	54%	64%	78%	73%	77%
13	Report said team included an evaluation specialist	340	13%	20%	16%	13%	7%	31%	5%	11%
14	Evaluation team included local members	340	30%	37%	30%	34%	22%	0%	29%	38%
15	Report indicated Conflict of Interest forms were signed	340	3%	0%	3%	4%	7%	0%	2%	8%
16	Study limitations were included	340	51%	37%	59%	53%	46%	61%	43%	35%
17	Report structured to respond to questions (not issues)	257	45%	56%	48%	38%	41%	60%	34%	52%
18	Evaluation questions addressed in report (not annexes)	232	62%	74%	61%	63%	50%	67%	60%	68%
19	Reason provided if some questions were not addressed	88	9%	0%	16%	6%	13%	0%	0%	0%
20	Social science methods (explicitly) were used	319	77%	77%	78%	72%	80%	92%	66%	84%

*The sequence of numbered factors in this table runs from 1–39, but actually includes only 37 numbered elements. Element 21 on the transparency between data and the source of that data and Element 36 on the inclusion of each and every data collection instrument were eliminated during the analysis phase of the meta-evaluation when an item analysis revealed that interrater reliability checks on these two items showed that they were not within acceptable limits.

META-EVALUATION OF QUALITY AND COVERAGE OF USAID EVALUATIONS 2009–12

Rating Factor		N =	Average All Regions	USAID Regions and USAID/W						
#	Description			AfPak (35)	AFR (128)	ASIA (55)	E&E (41)	LAC (42)	ME (26)	USAID/W (13)
22	Findings supported by data from range of methods	331	74%	82%	72%	79%	72%	77%	67%	76%
23	Findings distinct from conclusions/recommendations	340	41%	49%	44%	40%	41%	15%	29%	54%
24	Findings are precise (not simply “some,” “many,” or “most”)	310	66%	81%	69%	72%	57%	33%	63%	60%
25	Unplanned/unanticipated results were addressed	340	15%	14%	19%	14%	12%	15%	9%	4%
26	Alternative possible causes were addressed	340	10%	9%	10%	11%	12%	15%	7%	4%
27	Evaluation findings sex disaggregated at all levels	274	20%	38%	26%	10%	6%	22%	18%	14%
28	Report discusses differential access/benefit for men/women	262	32%	69%	29%	34%	17%	20%	31%	20%
29	Recommendations—not full of findings, repetition	318	59%	56%	65%	65%	53%	50v	58%	38%
30	Recommendations—specific about what is to be done	318	72%	71%	69%	78%	76%	75%	76%	69%
31	Recommendations—specify who should take action	318	49%	53%	47%	41%	58%	58%	54%	50%
32	Recommendations—clearly supported by findings	318	80%	88%	81%	82%	87%	83%	65%	65%
33	SOW is included as a report Annex	340	58%	63%	55%	64%	56%	69%	48%	65%
34	Annex included list of sources	340	78%	63%	82%	80%	88%	77%	69%	81%
35	Annex included data collection instruments	340	61%	60%	65%	62%	56%	77%	50%	58%
37	Statements of differences included as an annex	340	4%	0%	2%	4%	12%	15%	5%	4%
38	Report explains how data will transfer to USAID	340	2%	6%	2%	2%	0%	0%	5%	0%
39	Evaluation SOW includes Evaluation Policy Appendix I	212	6%	0%	4%	5%	11%	10%	8%	6%
N/A	Number of evaluation questions was 10 or fewer	206	43%	32%	32%	50%	41%	57%	52%	70%

Percentage of Evaluation by Sector Rated Positively on Evaluation Quality Factors

Table 64 below shows the percentages of evaluations between 2009 and 2012 by sector that were rated positively on each of 37 quality factors and an unnumbered factor on whether evaluation reports addressed 10 or fewer evaluation questions. As with the table above, which showed these percentages by region, the number of evaluations by sector is shown at the top of the column. Here again a caveat should be noted, that percentages are only percentages of the total for a sector when the number of data points in Column 3 equals 340. When Column 3 is a lower figure, it is likely that fewer than the total number of evaluations for any given sector were rated on that factor.

Table 64. Quality Factor Ratings By Sector
(among those evaluations on which data on a factor were available)

Rating Factor		N =	Average All Sectors	USAID Evaluation Sectors/Topics					
#	Description			AG (48)	DG (77)	ED (27)	EG (68)	Health (100)	Other (20)
1	Executive summary mirrors report all critical elements	323	45%	38%	43%	60%	42%	48%	56%
2	Project characteristics described	340	90%	90%	87%	93%	85%	96%	80%
3	Project “theory of change” described	340	74%	79%	77%	70%	68%	78%	65%
4	Management purpose described	314	80%	89%	76%	74%	79%	74%	79%
5	Questions were linked to purpose	190	99%	96%	100%	100%	98%	100%	100%
6	Questions in report same as in SOW	121	50%	44%	51%	40%	41%	56%	56%
7	Written approval for changes in questions obtained	62	6%	0%	0%	17%	27%	0%	0%
8	Data collection methods described	340	90%	94%	84%	96%	84%	95%	90%
9	Data collections methods linked to questions	265	23%	21%	16%	30%	19%	24%	44%
10	Data analysis method described	340	33%	40%	26%	52%	28%	30%	45%
11	Data analysis methods linked to questions	105	31%	18%	20%	58%	42%	21%	56%
12	External team leader	281	71%	77%	75%	81%	56%	72%	73%
13	Report said team included an evaluation specialist	340	13%	15%	14%	11%	16%	11%	15%
14	Evaluation team included local members	340	30%	31%	34%	30%	23%	31%	25%
15	Report indicated Conflict of Interest forms were signed	340	3%	0%	6%	11%	0%	4%	0%
16	Study limitations were included	340	51%	65%	48%	52%	48%	47%	50%
17	Report structured to respond to questions (not issues)	257	45%	46%	50%	43%	47%	40%	53%
18	Evaluation questions addressed in report (not annexes)	232	62%	73%	54%	79%	57%	62%	64%
19	Reason provided if some questions were not addressed	88	9%	0%	11%	50%	5%	9%	0%
20	Social science methods (explicitly) were used	319	77%	76%	73%	81%	80%	75%	83%
22	Findings supported by data from range of methods	331	74%	79%	65%	81%	77%	74%	70%
23	Findings distinct from conclusions/recommendations	340	41%	42%	34%	56%	46%	38%	50%
24	Findings are precise (not simply “some,” “many,” or “most”)	310	66%	78%	48%	70%	73%	65%	78%

Rating Factor		N =	Average All Sectors	USAID Evaluation Sectors/Topics					
#	Description			AG (48)	DG (77)	ED (27)	EG (68)	Health (100)	Other (20)
25	Unplanned/unanticipated results were addressed	340	15%	12%	22%	22%	10%	11%	15%
26	Alternative possible causes were addressed	340	10%	15%	12%	4%	10%	8%	5%
27	Evaluation findings sex disaggregated at all levels	274	20%	27%	15%	40%	13%	16%	37%
28	Report discusses differential access/benefit for men/women	262	32%	48%	34%	30%	25%	23%	50%
29	Recommendations—not full of findings, repetition	318	59%	66%	59%	56%	47%	68%	42%
30	Recommendations—specific about what is to be done	318	72%	68%	71%	74%	74%	73%	74%
31	Recommendations—specify who should take action	318	49%	52%	54%	56%	43%	44%	63%
32	Recommendations—clearly supported by findings	318	80%	82%	84%	81%	71%	79%	84%
33	SOW is included as a report Annex	340	58%	56%	65%	41%	56%	57%	65%
34	Annex included list of sources	340	78%	69%	80%	67%	81%	82%	85%
35	Annex included data collection instruments	340	61%	60%	54%	63%	57%	67%	65%
37	Statements of differences included as an annex	340	4%	2%	6%	0%	3%	5%	10%
38	Report explains how data will transfer to USAID	340	2%	4%	1%	4%	1%	1%	10%
39	Evaluation SOW includes Evaluation Policy Appendix I	212	6%	6%	8%	15%	5%	3%	0%
N/A	Number of evaluation questions were 10 or fewer	206	43%	44%	44%	53%	40%	44%	36%

Percentage of USAID Forward Evaluations Rated Positively on Quality Factors Compared With Non-USAID Forward Evaluations for the Same Period

The final summary table under Question I focuses on a comparison between evaluations completed, for the most part, after July 2011. Evaluations designated as USAID Forward evaluations reflect a special effort on the part of USAID Missions to produce high-quality evaluations in line with new standards established in USAID's 2011 Evaluation Policy. Missions then nominated evaluations as USAID Forward evaluations and PPL/LER reviewed them against 10 evaluation quality criteria. At the start of the meta-evaluation in early 2013, PPL/LER shared with MSI a listing of the USAID Forward evaluations received and reviewed. Table 65 below shows the number of USAID Forward evaluations identified by PPL/LER and that fell within MSI's sample, disaggregated by bureau. The table also shows the number of evaluations from each bureau for the same time period (July 2011 to December 2012) in the meta-evaluation study sample. These 154 evaluations, of which 45 percent were USAID Forward evaluations, serve as a basis for comparison of the frequency with which evaluations in each group complied with USAID quality expectations on the 37 quality factors in the checklist, plus an unnumbered question that tracked whether the number of evaluation questions addressed was 10 or fewer.

Table 65. Presence of USAID Forward Evaluations By Region, July 2011 Through December 2012

USAID Evaluations Between July 2011 and December 2012	USAID REGIONS and USAID/W							Total
	AFPAK	AFR	ASIA	E&E	LAC	ME	USAID /W	
USAID Forward Evaluations	10	14	15	10	11	9	0	69
Non-USAID Forward Evaluations	9	36	8	11	11	4	6	85
Total	19	50	23	21	22	13	6	154

Table 66 below compares the percentage of evaluations in USAID Forward and non-USAID Forward clusters defined in Table 66 that were scored positively on the meta-evaluation's quality rating factors. In Table 66, Column 2 shows this percentage for USAID Forward evaluations while Column 3 shows the parallel percentage for non-USAID Forward evaluations. Column 5 shows the difference between these groups of evaluations on each factor. The table is organized so that the difference between these groups is shown in descending order in Column 5. Divider rows in the table indicate the factors in which USAID Forward evaluations were more compliant with USAID guidance on quality factors than non-USAID Forward evaluations, where they were equal on a particular factor, and where non-USAID Forward evaluations were more consistent with USAID evaluation standards. Of the 37 factors rated, USAID Forward evaluations outperformed non-USAID Forward evaluations on 22 quality factors (58 percent) while non-USAID Forward evaluations did better at complying with USAID evaluation quality guidance on 13 factors (34 percent). On three factors they were equal.

Table 66. Ratings of USAID Forward Evaluations Compared With Non-USAID Forward Evaluations

Evaluation Report Quality Factors (Full List)		USAID Forward Evaluations Percentage = Yes (2011–12) (N = 69)	All Other USAID Evaluations Percentage = Yes (2011–12) (N = 85)	Difference Between USAID Forward Evaluations and Others *
#	Description			
USAID Forward Ratings Exceed Non-USAID Forward Ratings				
6	Questions in report same as in SOW	74%	44%	30%
24	Findings are precise (not simply “some,” “many,” or “most”)	71%	58%	13%
4	Management purpose described	90%	77%	13%
12	External team leader	78%	66%	12%
33	SOW is included as a report annex	81%	69%	12%
17	Report structured to respond to questions (not issues)	48%	36%	12%
22	Findings supported by data from range of methods	82%	71%	11%

* In percentage points.

META-EVALUATION OF QUALITY AND COVERAGE OF USAID EVALUATIONS 2009–12

Evaluation Report Quality Factors (Full List)		USAID Forward Evaluations Percentage = Yes (2011–12) (N = 69)	All Other USAID Evaluations Percentage = Yes (2011–12) (N = 85)	Difference Between USAID Forward Evaluations and Others *
14	Evaluation team included local members	38%	27%	11%
35	Annex included data collection instruments	77%	67%	10%
15	Report indicated Conflict of Interest forms were signed	13%	4%	9%
13	Report said team included an evaluation specialist	19%	11%	8%
18	Evaluation questions addressed in report (not annexes)	66%	58%	8%
1	Executive summary mirrors report in all critical elements	54%	49%	5%
3	Project “theory of change” described	80%	75%	5%
34	Annex included list of sources	86%	81%	5%
2	Project characteristics described	94%	91%	3%
16	Study limitations were included	64%	61%	3%
28	Report discusses differential access/benefit for men/women	32%	30%	2%
20	Social science methods (explicitly) were used	82%	80%	2%
38	Report explains how data will transfer to USAID	6%	4%	2%
32	Recommendations—clearly supported by findings	81%	79%	2%
11	Data analysis methods linked to questions	25%	23%	2%
No Difference				
7	Written approval for changes in questions obtained	9%	9%	0%
8	Data collection methods described	94%	94%	0%
37	Statements of differences included as an annex	6%	6%	0%
USAID Forward Ratings Lower Than Non–USAID Forward Ratings				
29	Recommendations—not full of findings, repetition	60%	61%	–1%
26	Alternative possible causes were addressed	8%	9%	–1%
23	Findings distinct from conclusions/recommendations	45%	46%	–1%
31	Recommendations—specify who should take action	50%	51%	–1%

Evaluation Report Quality Factors (Full List)		USAID Forward Evaluations Percentage = Yes (2011–12) (N = 69)	All Other USAID Evaluations Percentage = Yes (2011–12) (N = 85)	Difference Between USAID Forward Evaluations and Others *
5	Questions were linked to purpose	98%	100%	–2%
N/A	Number of evaluation questions was 10 or fewer	26%	29%	–3%
39	Evaluation SOW includes Evaluation Policy Appendix I	9%	11%	–2%
10	Data analysis method described	35%	38%	–3%
9	Data collections methods linked to questions	22%	26%	–4%
25	Unplanned/unanticipated results were addressed	12%	18%	–6%
27	Evaluation findings sex disaggregated at all levels	19%	25%	–6%
19	Reason provided if some questions were not addressed	0%	7%	–7%
30	Recommendations—are specific about what is to be done	68%	82%	–14%

Question 2: At this point in time, on which evaluation quality aspects or factors do USAID’s evaluation reports excel and where are they falling short?

Under this question, MSI clustered evaluation rating data to identify those quality factors in which USAID evaluations already performed well and those for which ratings were low. In addition to reviewing these strengths and weaknesses, this section discusses information obtained through small group discussions and team leader surveys about USAID efforts to ensure or strengthen evaluation quality by improving evaluation management practices, including evaluation SOWs and quality control efforts. The data indicate that these methods are being used by some but not necessarily all USAID Missions and offices.

Strong and Weak Evaluation Quality Factor Ratings

To address Question 2, MSI used data collected on 37 quality factors included in the evaluation review checklist (Annex C).^{*} Data on one additional quality factor—the number of evaluations that addressed 10 or fewer questions—was also used to answer this question since USAID guidance has long suggested that a small number of evaluation questions play an important role in evaluation quality. This addition increased the number of quality factors that were explored to identify where evaluation quality is currently strong and where there is room for improvement to 38.

To identify, in terms of the evaluation quality factors mentioned, where USAID evaluations are currently strongest and where weaknesses exist, MSI organized the evaluation quality factors identified above into

^{*}As noted earlier, 2 of the 39 evaluation quality factors on the checklist instrument were deemed by MSI to have low interrater reliability, making information on these factors unreliable.

four clusters reflecting the frequency with which each factor was rated positively in MSI’s review process. The four clusters, shown in Table 67, are not equal. The top cluster deliberately sets a fairly high bar for good performance, which is consistent with USAID initiatives that focus on standards for and improvements in evaluation quality. While the number of factors in the top cluster is high, data from the meta-evaluation indicates that USAID evaluations are already meeting these criteria for roughly a quarter of the evaluation quality factors in the meta-evaluation’s quality factors checklist.

Table 67. Evaluation Quality Factor Rating Clusters

Good	80% or more evaluations were rated positively on each quality factor in this cluster
Fair	50% to 79% of evaluations were rated positively on each quality factor in this cluster
Marginal	25% to 49% of evaluations were rated positively on each quality factor in this cluster
Weak	Fewer than 25% of evaluations were rated positively on each quality factor in this cluster

To address meta-evaluation Question 2, MSI applied the ratings for the 37 quality factors identified above to evaluations completed in 2012, the final year of the meta-evaluation study period. The final year of the study period was chosen because it is the most effective in indicating USAID’s current status in terms of evaluation quality. Using this data also serves as a natural baseline for assessing future performance. An added benefit is that 2012 was also the year with the highest percentage of positive ratings for many, though not all, of the quality factors assessed by the checklist.

Table 68 below applies the cluster definitions to the evaluation quality factors themselves. Data included in this table only reflects USAID evaluations completed in 2012 for the reasons stated above. This table also provides information on where guidance on each evaluation quality factor has appeared in recent years, including USAID’s 2011 Evaluation Policy. These references are included to indicate how familiar USAID staff and evaluation teams could or should be with respect to each quality factor.

As Table 68 shows, ratings on evaluation quality factors for 2012 evaluations are fairly evenly distributed across the clusters. Of the 37 factors examined, 24 percent were rated at the “Good” performance level, 26 percent achieved the “Fair” ratings level, 18 percent were rated at the “Marginal” level, and 32 percent received “Weak” ratings.

Particular attention should be paid to the way in which 2012 evaluations performed against evaluation quality factors formally introduced into USAID guidance through the 2011 Evaluation Policy. These factors are identified in red in Table 68.

- Two of these “new” evaluation factors received ratings in the top or “Good” cluster for 2012 evaluations. These two factors were the indication of external team leaders and the use of social science methods to conduct evaluations.
- The other five evaluation quality standards introduced in the 2011 Evaluation Policy were not found to be widely applied across the 2012 evaluations, and therefore were included in the “Weak” cluster.

Table 68. Quality Factors Clustered into Four Performance Levels

Evaluation Report Quality Factors (Full List)		2012 Evaluations Percent = “Yes” on Quality Factor	Earlier USAID Guidance on Importance and Application of this Evaluation Quality Factor	Quality Factor Introduced or Reinforced in 2011 Evaluation Policy
#	Description			
Good—80 Percent or More Scored Positively				
2	Project characteristics described	91%	2006 Reports* & 2008 ADS	
4	Management purpose described	81%	2008 ADS & 2010 TIPS [†]	●
5	Questions were linked to purpose	99%	2008 ADS & 2010 TIPS	●
8	Data collection methods described	95%	2008 ADS	●
12	External team leader	83%		New in 2011 Policy
20	Social science methods (explicitly) were used	84%		New in 2011 Policy
22	Findings supported by data from a range of methods	80%	2006 Reports	●
34	Annex included list of sources	83%	2010 TIPS on Reports	●
35	Annex included data collection instruments	81%	2010 TIPS on Reports	●
Fair—Between 50 Percent and 79 Percent Scored Positively				
32	Recommendations—clearly supported by findings	79%	2010 TIPS	●
3	Project “theory of change” described	74%	2006 Reports, 2010 TIPS	
6	Questions in report same as in SOW	74%	2006 Reports	●
16	Study limitations were included	62%	2006 Reports, 2010 TIPS	●
18	Evaluation questions addressed in report (not annexes)	74%	2006 Reports	●
24	Findings are precise (not simply “some,” “many,” or “most”)	67%	2010 TIPS	
29	Recommendations—not full of findings, repetition	64%	2008 ADS and later	●
30	Recommendations—specific about what is to be done	77%	2010 TIPS	●
33	SOW is included as a report annex	74%	2010 TIPS	●

*This reference is to *Constructing an Evaluation Report* (2006), which was prepared for USAID by Richard Blue and Molly Hageboeck, MSI.

†TIPS 2010 refers to the USAID Tips summary on Constructing Evaluation Reports, which summarized and updated the 2006 report with the same title.

META-EVALUATION OF QUALITY AND COVERAGE OF USAID EVALUATIONS 2009–12

Evaluation Report Quality Factors (Full List)		2012 Evaluations Percent = “Yes” on Quality Factor	Earlier USAID Guidance on Importance and Application of this Evaluation Quality Factor	Quality Factor Introduced or Reinforced in 2011 Evaluation Policy
#	Description			
N/A	Number of evaluation questions was 10 or fewer	52%	2008 ADS—“a small number”	
17	Report structured to respond to questions (not issues)	51%	2006 Reports, 2010 TIPS	
Marginal—Between 25% and 49% Scored Positively				
1	Executive summary mirrors critical report elements	45%	2008 ADS & 2010 TIPS	
10	Data analysis method described	34%	2008 ADS (SOW)	●
14	Evaluation team included local members	35%	2008 ADS	●
23	Findings distinct from conclusions/recommendations	48%	2006 Reports, 2010 TIPS, 2008 ADS	●
28	Report discusses differential access/benefit for men/women	40%	2008 ADS	●
31	Recommendations—specify who should take action	45%	2006 Reports, 2010 TIPS	●
Weak—Less Than 25% Scored Positively				
9	Data collections methods linked to questions	22%	2006 Reports, SOWs—Good Practice (2011),* How-To 2012	
11	Data analysis methods linked to questions	19%	2006 Reports, How-To 2012	
13	Report said Team included an Evaluation Specialist	19%	2008 ADS	●
15	Report indicated Conflict of Interest forms were signed	12%		New in 2011 Policy
7	Written approval for changes in questions obtained	12%		New in 2011 Policy
19	Reason provided if some questions were not addressed	10%		New in 2011 Policy
25	Unplanned/unanticipated results were addressed	14%		
26	Alternative possible causes were addressed	10%		
27	Evaluation findings sex disaggregated at all levels	22%	2008 ADS thru current	●
37	Statement of differences included as an annex	7%	2006 Reports, 2010 TIPS	●
38	Report explains how data will transfer to USAID	5%		New in 2011 Policy
39	Evaluation SOW includes Evaluation Policy Appendix I	8%		New in 2011 Policy

*SOWs—Good Practice refers to Evaluation Statements of Work; Good Practice Examples (2011) prepared for USAID by Micah Frumkin and Emily Kearney, MSI.

Beyond those evaluation quality factors identified as new, there are a reasonably large number of USAID guidance materials that were available before the start of the meta-evaluation study period.

In the “Good” cluster, guidance on five quality factors has been available since 2008 or earlier. USAID guidance for two other factors—the inclusion of lists of sources and copies of instruments—appeared by 2010, and are treated as standard practice in evaluation literature and in formal evaluation training courses provided by USAID for its staff since 2000.

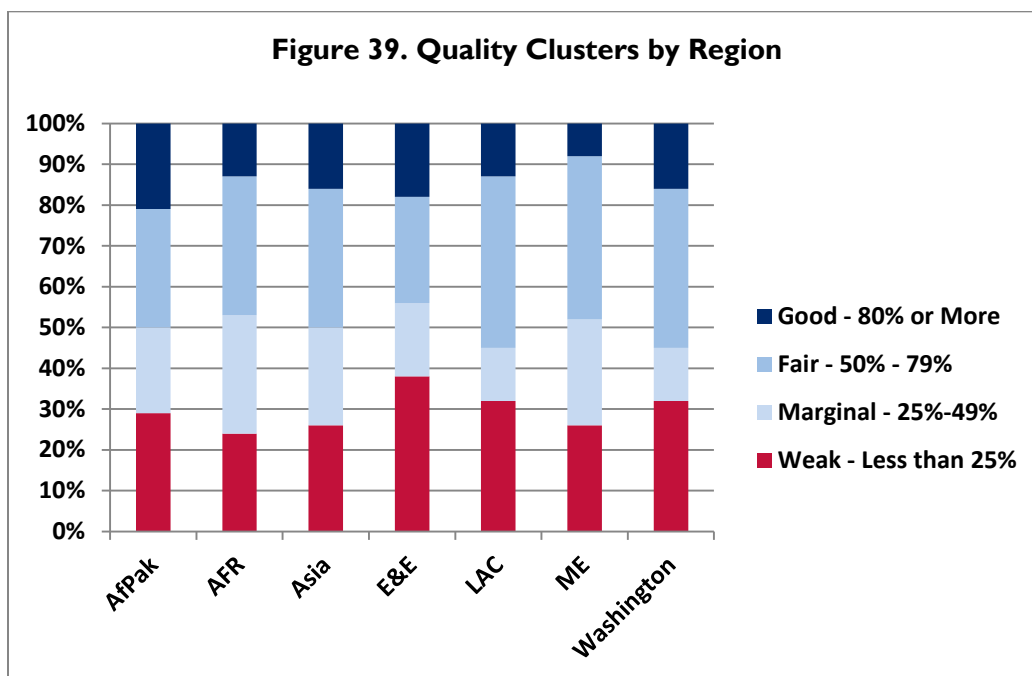
In other words, all of the evaluation quality factors rated as being very widely applied in 2012 and included in the “Good” cluster, are factors with which USAID staff and evaluators who routinely work on USAID evaluations have long been familiar. The same can be said for many evaluation quality factors that fall into the “Fair” and “Marginal” clusters. Thus, it is unlikely that familiarity with a requirement, taken alone, distinguishes between factors that scored “Fair” or “Marginal” rather than “Good.”

Among factors clustered in the “Weak” group that were not newly introduced through the 2011 Evaluation Policy, two quality factors stand out as being particularly problematic. The first of these is the low rate at which 2012 evaluation reports noted the presence of an evaluation specialist on the team. The other is the low frequency with which evaluations were rated as including sex-disaggregated data at all levels in which the evaluation focused. Given that both are long standing USAID evaluation standards, the fact that they are not being applied at the same level as many other long standing evaluation guidelines suggests that other factors or issues may play a role. For example, in group interviews and the evaluation team leader survey conducted as part of the meta-evaluation, individuals familiar with USAID evaluation practices told MSI that USAID has a strong preference for sector specialists on evaluation teams. Additional data on this issue and others are discussed under Question I, which looks more closely at how this practice has changed, or fail to change, over the meta-evaluation study period.

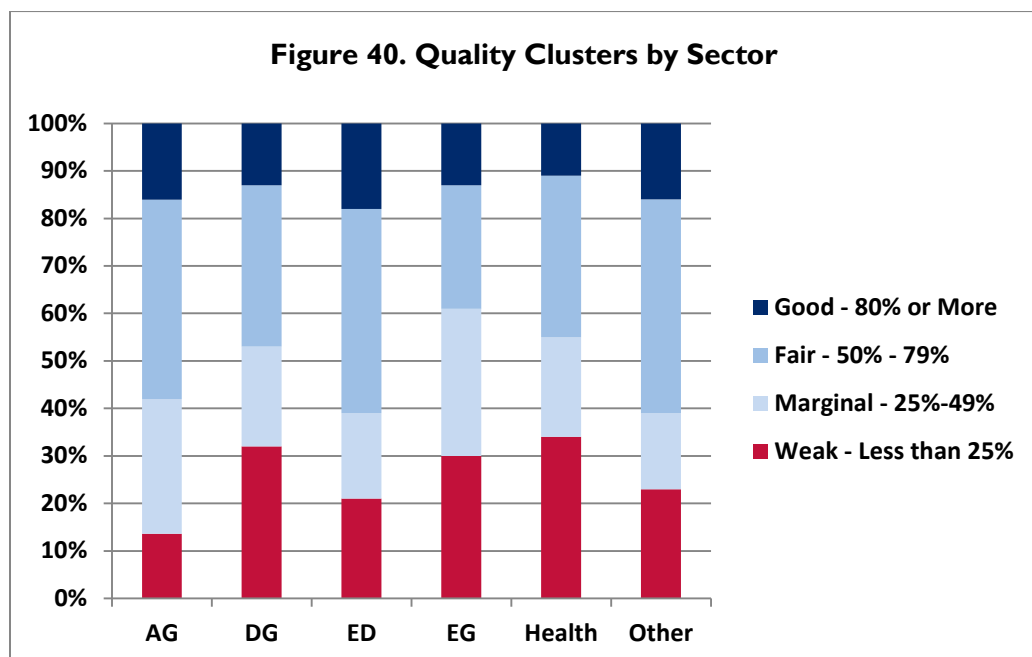
Setting aside the somewhat unique nature of several of the evaluation quality factors that fell into the “Weak” cluster, MSI’s review of factors in the “Fair” and “Marginal” clusters suggest that several of them involve simple oversights that might have been corrected as late in the evaluation process as USAID’s review of a draft report (e.g., making recommendations more specific, including the SOW as a report annex, describing the project’s “theory of change,” answering evaluation questions in the body of the report instead of in an annex, and the absence of a statement of study limitations).

In addition to examining how evaluations were rated on factors along the “Good” to “Weak” compliance spectrum, MSI also examined this distribution by region and sector. In the two figures below, the percentage of quality rating factors for which 80 percent or more of evaluations were scored positively is shown at the top with the lowest rated factors shown at the bottom. A useful way of comparing regions or sectors with one another is to follow the 50 percent line across the graph. This divides the percentage of rating factors that scored “Good” or “Fair” from lower clusters. Another focal point in this chart is at the bottom where factors in which fewer than 25 percent of evaluations received a positive score are highlighted in red.

Figure 39 focuses on geographic regions. USAID/W technical bureaus are included as a separate cluster. In two of the seven regions depicted, more than 50 percent of the quality factors were rated “fair” or better, these two were LAC and USAID/W. Other regions had a higher proportion of quality factors rated “Marginal” or “Weak.”

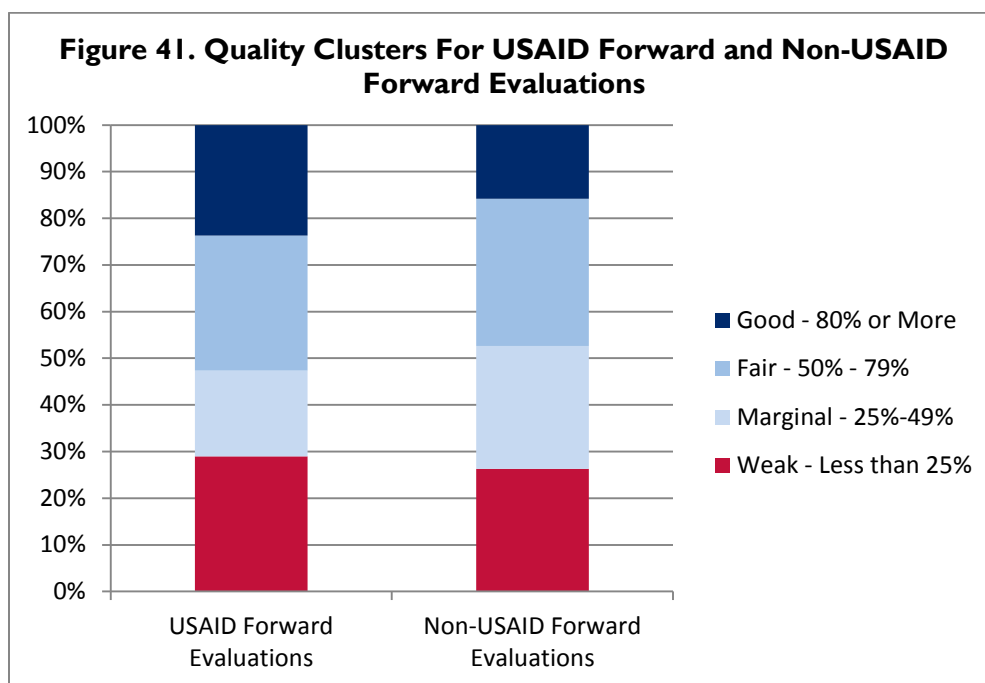


In Figure 40, the graph shows how evaluations rated in evaluation quality factor clusters by sector. In this graph, ratings on more than half of the quality rating factors fell in the “Fair” or “Good” category in three sectors: agriculture, education, and a small group of evaluations clustered under “other.” In this context, “other” is a category that included a small number of projects focused on communications, management, and other issues that fell outside the five main sectors considered by the meta-evaluation.



In addition to region and sector, MSI also examined USAID Forward evaluations and non-USAID Forward evaluations completed between July 2011 and December 2012 to see where these evaluations fell in relation to the four categories devised to assess quality strengths and weaknesses. As Figure 41 shows, USAID Forward evaluations had a higher percentage of quality factors on which 80 percent or

more were rated positively, thus placing them in the “Good” cluster. Also, when looking at the 50 percent breakpoint on the vertical axis, it is clear that USAID Forward evaluations had more factors that fell into the two top clusters (“Good” and “Fair”) than did non-USAID Forward evaluations. The differences between these two clusters of evaluations was not very substantial as both USAID Forward and non-USAID Forward evaluations had roughly the same number of evaluation quality factors in the lowest two clusters (“Marginal” and “Weak”).



USAID Evaluation Management Practices and Evaluation Quality

Data from the meta-evaluation indicated that some USAID staff are focusing on evaluation management practices as a way to improve evaluation quality. This section presents findings on evaluation SOW development and other steps in the evaluation process, which were obtained through small group interviews and reported on through the team leader survey. USAID attention paid to evaluation quality through a deliberate quality improvement effort under USAID Forward is also discussed here.

A. Evaluation Statements of Work (SOWs) as an Evaluation Quality Determinant

In small group interviews with USAID regional and technical bureau staff and firms that carry out evaluations for USAID, a number of observations were offered on evaluation SOWs as a factor affecting evaluation quality.

Overall, six of the firms that participated in these interviews said that the overall quality of USAID evaluation SOWs has improved—they are clearer and generally better. One technical office representative offered a similar perception. Three other firms, however, disagreed and said that the evaluation SOWs they have seen recently are either roughly the same as in the past or possibly worse. One technical representative pointed out that there is a difference between SOW quality in the field and USAID/W, and that improvements are taking place more so in Washington than in the field.

Among firms that felt evaluation SOWs have been improving, two indicated that USAID’s evaluation training courses are having a positive influence on the quality of evaluation SOWs. Among those

disagreeing, one firm opined that these trainings are insufficient and that SOW writers need to experience evaluations in the field to be able to write good ones. On this same point, five of the firms noted that the language being used in SOWs is often cut and pasted from the Evaluation Policy or include “buzz words” that writers have heard, but do not understand. This is demonstrated by the use of incorrect meanings for some terms or an apparent lack of understanding of the methodological implications of some of the Evaluation Policy’s guidance. This, they said, leads to poor quality SOWs and creates confusion in the firms about how to respond to such solicitations, both technically and financially.



USAID participant presenting exercise results during the USAID Evaluation for Evaluation Specialists (EES) training course. Washington, D.C., January 7-18, 2013. Photo Credit: Social Impact

Participants in small group discussions also commented on USAID’s process for reviewing SOWs before releasing them. Regarding the SOW peer-review process in the first year following USAID’s release of the 2011 Evaluation Policy, a regional representative stated that the review process is generally accepted and appreciated, though there was much resistance at first. One regional bureau representative stated that the SOW review process has increased SOW quality, although another regional representative said that, in their view, SOW quality may have decreased as a result of reviews. Four regional and two technical office staff representatives stated that the current process for reviewing SOWs does not provide enough direction on how, when, or by whom they should be done, leading to inconsistency in quality of reviews and reviews being done by people without enough evaluation knowledge.

B. USAID Evaluation Management/Quality Control Process

In addition to contributions USAID makes to evaluation quality through decisions reflected in evaluation SOWs, there are a number of activities which impact evaluation quality that evaluation managers and other USAID staff engage in after an evaluation team has been selected. In USAID’s evaluation courses, four of these activities are treated as evaluation quality control opportunities, and all of them were commented upon in data collected for the meta-evaluation through small group interviews and the team leader survey. When required, each of these quality control activities tend to be stated as evaluation deliverables in SOWs, and are discussed below.

1. A Required Summary of What is Known from Existing Documents and What Gaps Remain

Some evaluation SOWs ask evaluation teams to produce an inception report that summarizes what they have learned from their review of existing documents and project or program performance indicators. Such inception reports are typically organized around each evaluation question in their SOW and include a summary of the data gaps that must be filled in order to answer those questions. In some cases, this analysis of existing documents is part of an evaluation design product a team produces.

Evaluation inception reports are required by many United Nations organizations and the World Bank as part of their evaluation processes. A succinct definition of the scope of an inception report from one

such agency is provided below. Note that there are two elements to this definition: a) a structured presentation on the findings of a desk study or document review and b) a detailed evaluation design prepared by the actual team for the evaluation.

An Inception Report summarizes the review of documentation ("desk review") undertaken by an evaluator mandated by UNODC and specifies the evaluation methodology determining thereby the exact focus and scope of the exercise, including the evaluation questions, the sampling strategy and the data collection instruments. Consequently, the evaluator is expected to deliver an *Inception Report* as one of the key deliverables, which is shared with the Project Manager and the Independent Evaluation Unit for comments.

**Evaluation Handbook
United Nations Office on Drugs and Crime**

The first element of an inception report helps to ensure that most of the field work period is dedicated to gathering data that is not already available.

- The Team Leader Perceptions Survey included a question about this requirement and revealed that 46 percent of respondents had been asked to complete this task during their most recent evaluation.
- In small group interviews, two firms affirmed that in recent evaluations they had been asked to provide "inception reports" that synthesize what is already known from existing documents.

2. Submission and Approval of a Detailed Version of the Evaluation Design and Instruments Prior to the Start of Field Work

The second checkpoint the evaluation course recommends is the submission of a detailed version of the team's evaluation design for approval prior to the start of field work. This quality control activity helps to ensure that adequate methodologies have been developed to address all evaluation questions and that appropriate sampling plans exist where needed. It also helps to ensure that professional data collection instruments have been developed for both interviews and structured observations, and that attention has been paid to pretest and translate them.

- Responses to the Team Leader Perceptions Survey indicated that 71 percent of respondents were asked to this during their most recent evaluation for USAID.

3. Debriefs After Field Work and Data Analysis End, But Before the Writing of the Draft Report Begins

The third evaluation quality checkpoint involves a post-field work briefing by the evaluation team on its findings, conclusions, and recommendations, which normally occurs after data analysis but before the team begins to write the findings, conclusions, and recommendations sections of its report. This step helps ensure that data were collected on all evaluation questions and that there is an evidence-based progression from findings to conclusions to recommendations.

- Responses from 23 of the 25 respondents to the Team Leader Perceptions Survey (92 percent) reported that they had been asked to provide this type of briefing for USAID as part of their most recent evaluation.
- One firm participating in the small group

Evaluation Report Quality

"Reports are not improving because of a lack of capacity both in Washington and in the field, there are a lot of M&E people in the field that have never done M&E before but are now running teams."

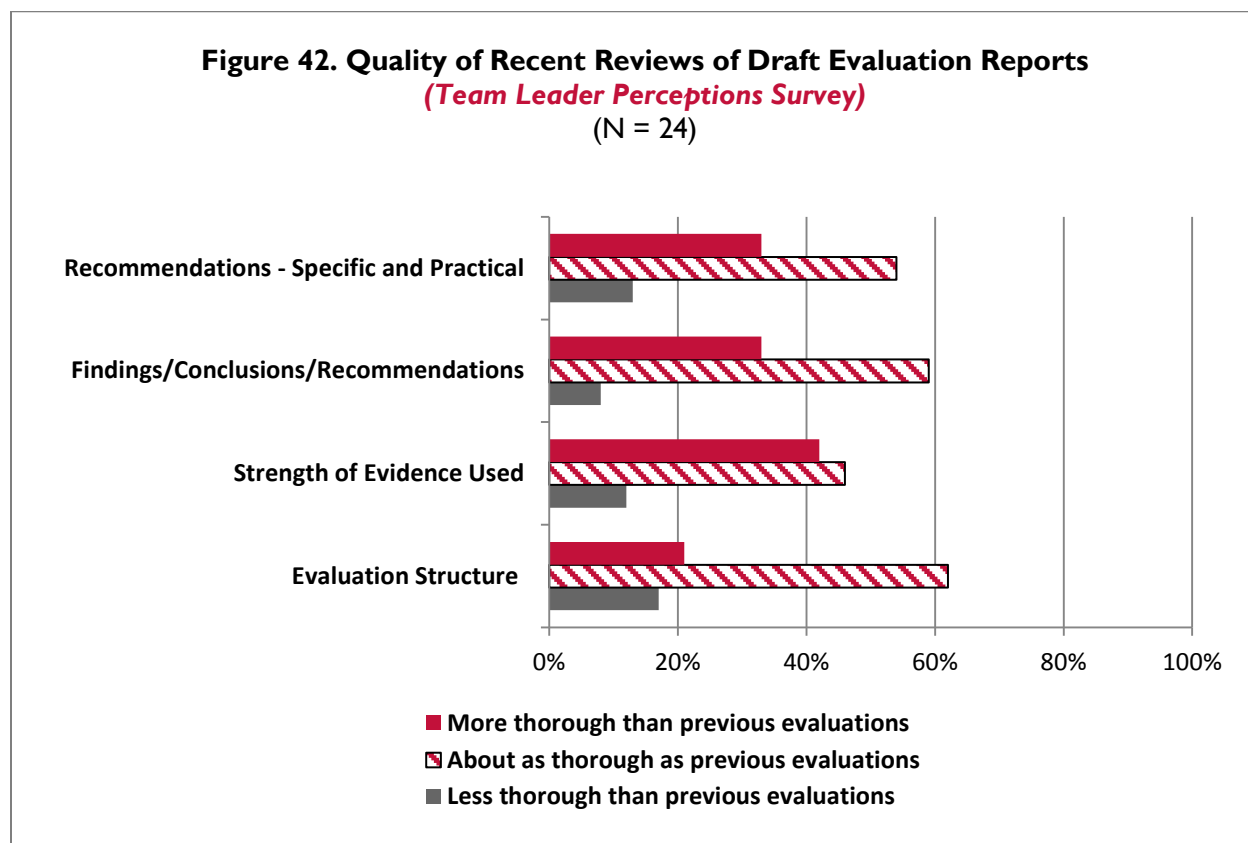
USAID Regional
Bureau Representative

interviews commented on having seen requests for reports on “findings to date” during the field data period of an evaluation.

4. USAID Review and Comments on Draft Reports and Acceptance of Final Evaluation Reports

The last quality control activity in this set is also the most familiar—USAID reviews of draft and final versions of evaluation reports. These reviews represent USAID’s final opportunity to ensure that evaluation reports are of the highest possible quality. While it is typically no longer possible at this point, due to a lack of resources, for changes in evaluation design, incorporation of new methods, or collection of additional data, there remain numerous opportunities for improving the structure and coverage of evaluation reports and their compliance with a wide range of USAID evaluation report guidelines.

In the meta-evaluation’s Team Leader Perceptions Survey, one question asked about the thoroughness with which recent evaluation reports have been reviewed compared with earlier evaluation reports that team leaders have been involved in. The team leaders’ responses, regarding the thoroughness of four aspects of USAID reviews of their evaluation reports, indicate that in most cases the quality of reviews remain unchanged. It also shows that some team leaders consider reviews of evaluation reports to have become more rigorous, particularly with respect to the strength of the evidence used to support evaluation conclusions and recommendations. Data on team leaders’ responses are in Figure 42.



Comments on evaluation report reviews were also offered in small groups meetings with USAID regional bureau staff, technical bureau staff, and firms that conduct evaluations. With respect to enhanced reviews of evaluation reports, one regional representative stated that report reviews are done inconsistently by Missions, with some sending them in to Washington and others not. This sentiment, with special reference to the timeliness of USAID evaluation report reviews, was echoed in a comment

received from one of several team leaders in the Team Leader Perceptions Survey who noted that timeliness is important:

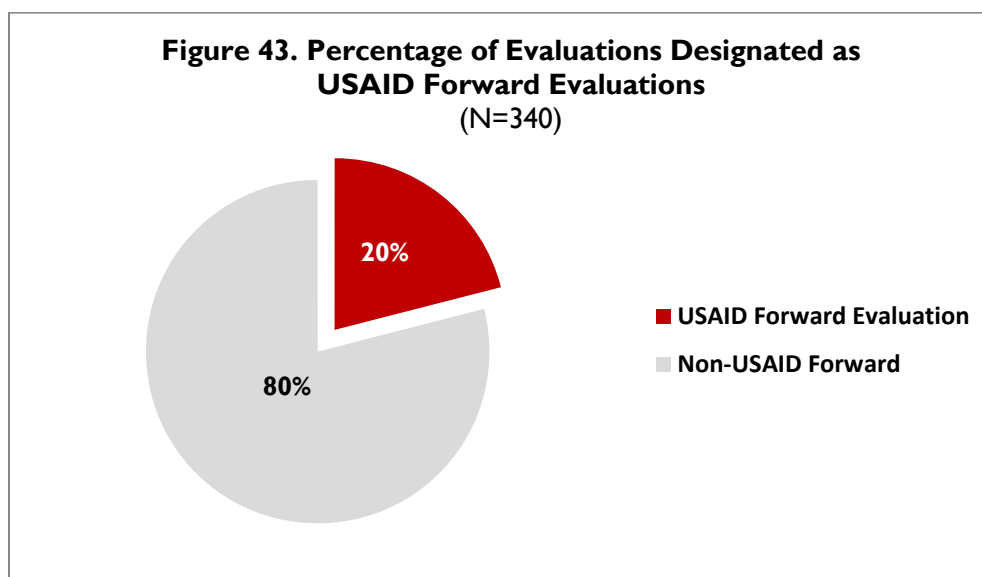
One USAID Mission took over two months to respond to the preliminary draft report of an evaluation and then nearly two months to respond to the revised report. This worked a serious hardship on the team leader and local consultants.

With respect to the value of the evaluation report review process, two firms stated that the review of reports is a good thing, but one of those firms also stated that it can become a bad thing when done too formulaically. Three evaluation firms stated that it is clear that USAID has been using checklists when reviewing evaluation reports. Another two firms noted that USAID appears to emphasize the linkages between findings, conclusions, and recommendations when reviewing evaluation reports.

Commenting more broadly on these kinds of quality control processes, three technical and regional representatives commented that the evaluation process at USAID is improving, while two technical and one regional technical bureau representative said that this process improvement has yet to translate to improvements in the quality of evaluation reports. Another regional bureau representative commented that there are now much timelier submissions of evaluation reports to the DEC though this does not directly reflect evaluation quality itself.

C. Designation as a USAID Forward Evaluation

Of the 340 evaluations included in this meta-evaluation, 69 (20 percent) are designated by PPL/LER as USAID Forward evaluations, as Figure 43 below illustrates. In addition to this meta-evaluation, USAID's evaluation office in PPL/LER uses other approaches for monitoring evaluation quality and encouraging regional and technical bureaus to undertake evaluations that employ the most rigorous possible methods to ensure that resulting evaluations are of high quality. Of note in this regard is its ongoing effort under USAID Forward to support Missions in achieving their annual targets for high-quality evaluations as defined by the standards set forth in USAID's 2011 Evaluation Policy.



Promoting High Quality Evaluations

"We introduced a new evaluation policy that's been called a "model for other federal agencies" by the American Evaluation Association. Today, third parties perform evaluations to a high-quality standard, and you can access all 186 [USAID Forward] evaluations from the last two years right now on your iPhone."

Dr. Rajiv Shah
USAID Administrator
USAID Forward Progress Meeting
March 20, 2013

In addition to designating evaluations as efforts to improve evaluation quality, PPL/LER rated these evaluations against a set of agreed upon criteria with the regional bureaus. Evaluations that met these criteria were then entered into a special database of USAID Forward evaluations on USAID’s website.

Question 3: What can be determined about the overall quality of USAID evaluation reports and where are the greatest opportunities for improvement?

In answering the two previous meta-evaluation questions, MSI drew on data extracted from evaluation reports using the basic evaluation characteristic description instrument (Annex C) and ratings given to each of 340 evaluations using the meta-evaluation’s quality checklist (Annex C). Findings presented under evaluation Question 1 also drew on the results of group discussions with USAID technical and regional bureau staff, results of group discussions with representatives of organizations that conduct evaluations for USAID, and on a survey of recent evaluation team leaders. To answer this third evaluation question, MSI needed a measure of “overall quality”, or a way of boiling down the meaning of all of the quality factor ratings discussed under meta-evaluation Questions 1 and 2 into a holistic measure of the quality of each evaluation. In addition to developing a single holistic quality measure, evaluation Question 3 implies the need for an approach to measuring overall quality that will help increase USAID’s understanding of what characteristics of evaluations—including both basic evaluation characteristics and ratings on individual quality factors—are most closely associated with overall evaluation quality.

MSI’s approach to developing an overall measure of evaluation quality or “evaluation quality score,” was informed by its review of previous USAID meta-evaluations. Historically, most meta-evaluations undertaken for USAID and other entities have used a checklist approach to rate evaluation report quality. Some organizations, such as UNICEF, use checklists to categorize evaluations into groups for future decision making while others use checklists to create a numeric score along the lines of the *Program Evaluation Model Meta-Evaluation Checklist* by Daniel L. Stufflebeam (1999), although there is little in the way of published information on applications of this latter type of meta-evaluation tool.

With one notable exception, previous USAID meta-evaluations have reported their findings on a factor-by-factor basis, in much the way MSI addressed meta-evaluation Questions 1 and 2. The exception MSI found was an evaluation “score” approach used in USAID’s 1983 meta-evaluation, conducted by the Triton Corporation. This meta-evaluation used data on a number of evaluation characteristics to create an overall “score” for each evaluation it analyzed. Triton then used this score to make comparisons by regions, sectors, and other factors of interest at the time, including whether evaluations were led by USAID staff or non-USAID personnel. While some of the volumes of the Triton report on this effort appear to have been lost over the years, those that have been scanned into the DEC indicate that Triton’s approach involved a 100 point scoring system constructed around nine evaluation quality factors, each of which had a set of subfactors.

Overall Evaluation Score Construction

Building on what could be recovered of the Triton model and taking into account changes over time in evaluation quality factors of interest, MSI constructed a composite evaluation quality score or index. The evaluation quality factors included in the overall “score” instrument (Annex C) were selected based on two criteria:

- First, the evaluation factors must have been in place prior to the start of the meta-evaluation’s four-year study period from 2009 through 2012. This requirement was intended to create a

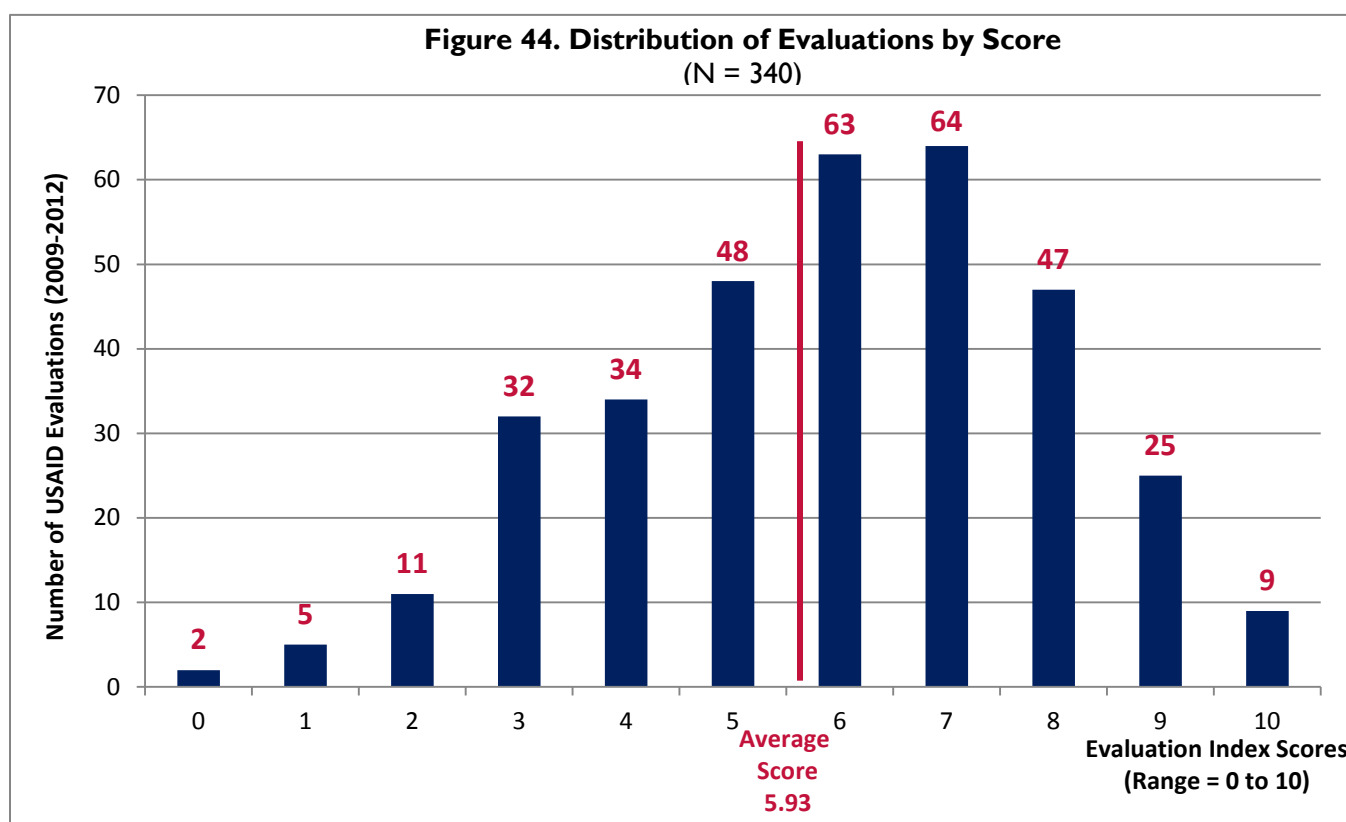
level playing field to ensure that both older and more recent evaluations, implemented under slightly different guidance, had a relatively equal chance to score well.

- Second, factors included in the scoring instrument have to seem important to evaluators as well as to USAID, irrespective of the type of evaluation, the geographic focus of the evaluation, or the sector that the evaluation focused on.

Using these criteria, and input from USAID staff, MSI selected 11 of the 37 evaluation quality factors from the evaluation quality checklist on which the meta-evaluation reported under Questions 1 and 2. These 11 factors were used to create a composite evaluation “score” sheet where the lowest possible score was zero and the highest possible score was 10 as two factors, background and theory of change, were merged into a single scoring point. This score sheet is provided in Annex C.*

Distribution of Overall Evaluation Scores

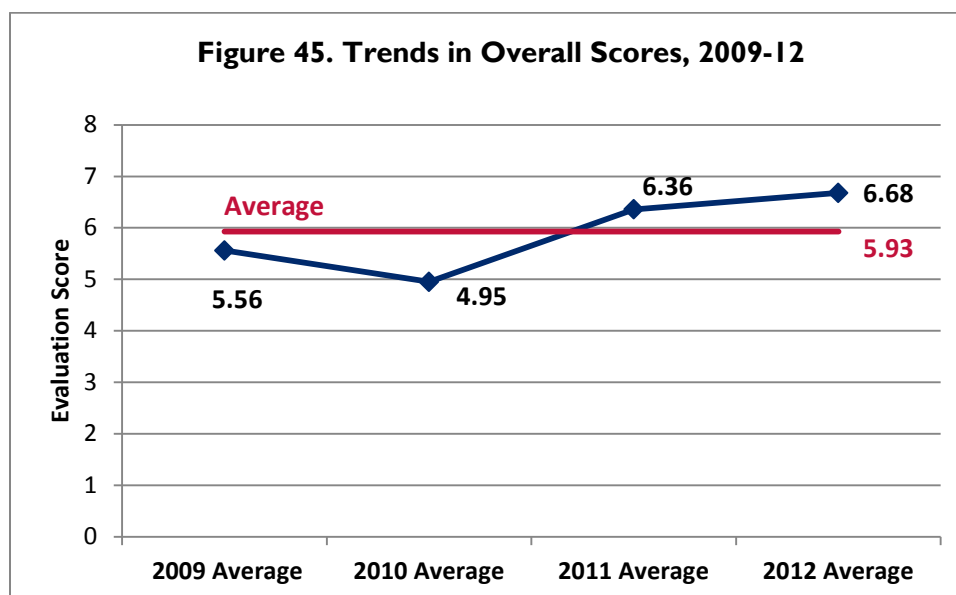
Figure 44 below displays the frequency with which evaluations completed between 2009 and 2012 received a score of between 0 and 10 on the simplified scoring system described above. The mean score for all evaluations included was 5.93, or slightly better than halfway to the top of the scoring system.



When calculated on an annual basis, average evaluation scores for 2009 through 2012 illustrate the path along which evaluation quality improved over this period, as shown in Figure 45. While this pattern of improvement over time was discussed under Question 1, on a factor-by-factor basis, the transformation

*Items in this 10 point scoring checklist include the following evaluation quality factors: 1, 2, and 3 combined; and 8, 10, 16, 20, 23, 32, 33, and 35. Items scored 1 if the rating on the item given by coders was a “yes” and 0 if the rating for the item had been “no.” Items were not “weighted” on an *a priori* basis when assigning a “score” to an evaluation. Instead, MSI used an item analysis to determine empirically which factors included in the score sheet were most highly associated with a high overall score.

of individual factor ratings into a composite evaluation score facilitates a determination of how important such differences are. A t-test comparing the average score for 2009 with the average score for 2012 indicates that the difference is statistically significant at the .05 level, indicating that this difference is not attributable to chance alone.

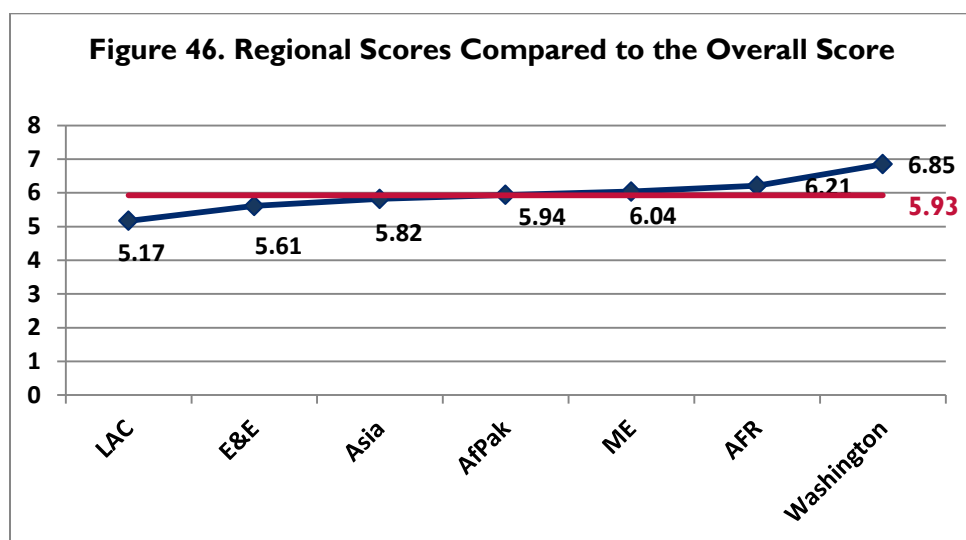


Transforming factor ratings into scores also makes it possible to roughly compare the quality of 2009–12 evaluations with those scored in 1983, as illustrated by the parallel tables of regional bureau scores in Table 69 below. While Table 69 ignores differences that invariably exist between the evaluation scores used in 1983 and those developed for this report in the period 2009 through 2012, two important points stand out. First, regional bureaus are only marginally different in terms of the quality of their evaluations, both then and now. Average evaluation quality scores in both cases fall just slightly above the halfway mark on the scoring range. In both eras, the distance between achieved scores on a regional basis and the top possible score appear about equal.

Table 69. Comparison of Average Evaluation Scores by Bureau, 1983 and 2009–12

USAID Evaluation Scores 1983		USAID Evaluation Scores 2009–12	
Bureau	Average Evaluation Score	Bureau	Average Evaluation Score
Impact CDIE	65.9	USAID/W	6.85
Asia	56.9	AFR	6.21
AFR	52.5	ME	6.04
USAID/W	52.1	AfPak	5.94
Near East	51.1	Asia	5.82
LAC	50.9	E&E	5.61
		LAC	5.17

On a regional basis, average scores for all 340 evaluations included in the meta-evaluation are shown in Figure 46, which displays the regional averages against a constant line represented by the average overall score for 2009–2012. The distance between the highest and lowest regional average evaluation score is 1.68 points on the 10 point scoring system used by MSI. This difference across seven geographic locations was not found to be statistically significant.

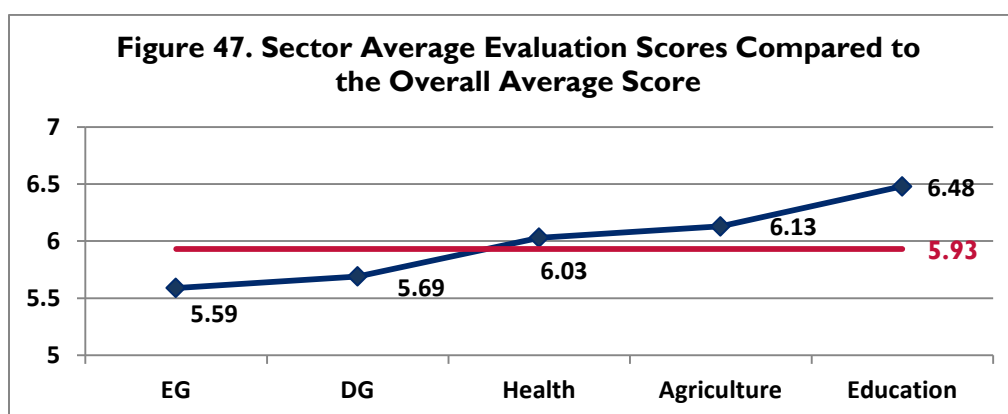


MSI also calculated average evaluation scores on a sector basis, as shown in Table 70.

Table 70. Average Scores by Sector

USAID Evaluation Scores 2009–12	
Sector	Average Evaluation Score
Education	6.48
Agriculture	6.13
Health	6.03
DG	5.69
EG	5.59

Figure 47 displays these sector averages compared with the overall average for 2009–12. The distance between the highest and lowest sector average evaluation score is .89 on a 10 point scoring scale, which is nearly half as wide a range as that found for regional bureaus. This difference across five sectors was, nevertheless, found to be statistically significant at the .05 level.



Another comparison MSI made was between evaluations identified as being USAID Forward evaluations, completed between July 2011 to December 2012, and other evaluations completed over that period as shown in Table 71. MSI compared the average evaluations scores for USAID Forward and non-USAID

Forward evaluations using a t-test and found that the difference between them was not statistically significant.

Table 71. Average Scores for USAID Forward and Other Evaluations

USAID Evaluation Scores 2011–12	
USAID Forward)	Average Evaluation Score
USAID Forward Evaluations	6.84
Non-USAID Forward Evaluations	6.33

Other Evaluation Factors Associated with Overall Evaluation Scores

In addition to comparing evaluation scores over time, by region, by sector, and for USAID Forward and non-USAID Forward evaluations, MSI analyzed evaluation scores in relation to other evaluation characteristics, including quality factor ratings on the meta-evaluation quality factor checklist. These comparisons fell into two groups: a) comparisons between overall scores for evaluations and factors which were not among the 10 used to construct the overall score and b) an item analysis in which the relationship between each of the 10 factors used to construct the “score” and the overall score.

A. Overall Evaluation Scores and Evaluation Factors/Ratings Not Used to Construct Scores

Using Chi Square tests, MSI examined a number of basic evaluation characteristics along with 27 evaluation checklist ratings that were not used to construct the “score” to determine which, if any, of these factors and ratings were associated with overall scores at a statistically significant level. Table 72 displays all of factors and ratings where a statistically significant association with overall score was found. These factors differed in terms of the strength of their association with an overall score, in that some were significant at the .05 level and others were significant at higher levels. Table 72 is organized to display the strongest associations between overall score and other factors and ratings in rank order based on the strength of association.

Table 72. Correlation Between Scores and Other Evaluation Characteristics

Factor #	Quality Rating Topic	Rank Order	Chi Square Value	Significance Levels*		
				.05	.01	.001
	Year in which evaluation was completed	1	29.601	●	●	●
25	Unplanned/unanticipated results were addressed	2	20.682	●	●	●
9	Data collections methods linked to questions	3	19.567	●	●	●
15	Report indicated Conflict of Interest forms were signed	4	18.196	●	●	●
22	Findings supported by data from range of methods	5	17.364	●	●	●
	Sector associated with program/project evaluated	6	17.330	●		
13	Report said team included an evaluation specialist	7	14.674	●	●	●
	Purpose was to support design of future strategies,	8	14.542	●	●	●

*For all numbered factors, the degrees of freedom in the Chi Square test was 2, but for some of the unnumbered items included in this table the degrees of freedom number was higher. This explains why Chi Square values that were roughly the same were not necessarily significant at all the same significance levels.

Factor #	Quality Rating Topic	Rank Order	Chi Square Value	Significance Levels*		
				.05	.01	.001
	programs, projects					
26	Alternative possible causes were addressed	10	13.720	●	●	
18	Evaluation questions addressed in report (not annexes)	11	13.346	●	●	
38	Report explains how data will transfer to USAID	12	11.984	●	●	
6	Questions in Report Same as in SOW	13	10.305	●		
17	Report structured to respond to questions (not issues)	14	10.193	●		
29	Recommendations—not full of findings, repetition	15	8.606	●		
14	Evaluation team included local members	16	7.344	●		

As Table 72 shows, the strongest association found between overall scores and other evaluation characteristics involved the year in which evaluations were completed. The fact that this factor has the strongest association with overall scores suggests that evaluation-related changes at USAID occurring over the 2009–12 time period played an important role in improving evaluation quality. Some of these evaluation changes at USAID can be seen in a timeline in Figure 1 at the front of this report.

Among the 15 unnumbered evaluation characteristics and numbered evaluation quality checklist factors found to be statistically associated with overall evaluation quality scores in Table 72, 11 appear to be reactive aspects of evaluation reports. This implies that some aspects—such as methods related to questions or unplanned results examined—are a function of some more dynamic variable that could explain why differences in quality are evidenced in the data set. Of the remaining four factors, two have already been discussed: 1) changes at USAID over time that may have triggered quality improvements, and 2) sector variations where evaluation experience and practices may differ in ways that affect quality (e.g., the education, agriculture, and health sectors may have better developed evaluation practices than USAID’s DG and EG sectors). The other two factors found to be statistically associated with overall evaluation scores and that could be potential drivers of higher quality are the presence of evaluation specialists on evaluation teams and the participation of local team members.

As Table 73 shows, the presence of an evaluation specialist on the team, as well as the presence of local team members, are both positively associated with average evaluation scores. The impact of having an evaluation specialist versus not having one (+1.09 score point difference) in terms of the difference in overall scores is about twice as strong as the presence of local team members (+.53 score point difference), but both are independently associated with score quality and hence warrant more detailed consideration of the factors their presence might affect.

Table 73. The Effect of Evaluation Specialists and Local Team Members on Scores

USAID Evaluation Scores 2012		USAID Evaluation Scores 2012	
Team Composition	Average Evaluation Score (Range 0–10)	Team Composition	Average Evaluation Score (Range 0–10)
Evaluation Specialist on Team	6.87	Local Team Members Involved	6.30
No Evaluation Specialist Identified	5.78	No Local Team Members Identified	5.77

To identify evaluation quality factors beyond the overall evaluation score that is associated with the presence of an evaluation specialist, MSI ran cross-tabulations for this variable with other evaluation variables and then ran Chi Square tests to identify important associations. As Table 74 shows, that analysis revealed that the presence of an evaluation specialist is associated, at a statistically significant level, with 10 meta-evaluation quality factors from the study's checklist of quality factors.

Most factors related to the reported presence of an evaluation specialist focus on evaluation design and methods. Notably, the percentage of evaluations that had an evaluation specialist on the team did not increase dramatically over the four years nor was a strong relationship found between the presence of an evaluation specialist and the sector on which an evaluation focused.

Table 74. Quality Factors Associated with the Presence of an Evaluation Specialist

Quality Factors Associated with an Evaluation Specialist Being Identified as a Team Member		Chi Square Value	Significance Levels	
#	Quality Rating Topic		.05	.01
14	Evaluation team included local members	10.492	●	●
35	Annex included data collection instruments	6.746	●	●
17	Report structured to respond to questions (not issues)	6.398	●	
9	Data collections methods linked to questions	5.938	●	
10	Data analysis method described	5.574	●	
23	Findings distinct from conclusions/recommendations	5.172	●	
12	External team leader	4.965	●	
33	SOW is included as a report annex	4.327	●	
15	Report indicated Conflict of Interest forms were signed	4.170	●	
11	Data collections methods linked to questions	3.951	●	

MSI repeated this type of analysis for local evaluation team members but found no statistically significant relationship between the presence of local team members and evaluation quality factors other than the presence of an evaluation specialist on the evaluation team, as shown in Table 75. The fact that the presence of local team members on evaluation teams was not, in and of itself, statistically associated with any specific evaluation quality scores would tend to rule this factor out as a driver of quality.

Table 75. Quality Factors Associated with the Presence of Local Evaluation Team Members

Quality Factors Associated with the Participation of Local Team Members		Chi Square Value	Significance Levels	
#	Quality Rating Topic		.05	.01
13	Evaluation Team Included Evaluation Specialists	10.492	●	●

At the same time, the presence of local team members is not likely to have been a result of identifying an evaluation specialist on a team. Rather, the presence of both of these types of individuals, as well as the presence of an external evaluation team leader, tends to be driven by requirements in USAID

evaluation SOWs. Thus, the presence of both local team members and an evaluation specialist may be better explained by the level of resources available for an evaluation, its duration, and the size and type of team USAID thought justified. Absent data on evaluation cost and duration, however, MSI was unable to test these hypotheses using meta-evaluation data.

In a third analysis of this type, MSI examined the quality factors associated with the indication of an external team leader. Indicating an external team leader has also been found to be linked to both overall quality scores and the presence of an evaluation specialist on the team, though quantitatively those cases were few in number. What Table 76 below shows is that the presence of an external evaluation team leader is associated with several important evaluation quality factors, even though it not found to be associated with overall evaluation quality at a statistically significant level. It is not surprising to find an association between evaluation team leaders and factors such as providing clear support for recommendations in the findings or ensuring that Conflict of Interest forms are signed, since these are often the responsibilities of an evaluation team leader. More surprising was the association between external team leaders and the presence of sex-disaggregated data, which was found in 22 percent of the 340 evaluations examined. No questions were asked in the team leader survey or small group interviews that would provide insights on why these two factors were found to be closely linked.

Table 76. Quality Factors Associated with the Presence of an External Team Leader

Quality Factors Associated with an Evaluation Specialist Being Identified as a Team Member		Chi Square Value	Significance Levels	
#	Quality Rating Topic		.05	.01
32	Recommendations—clearly supported by findings	5.965	●	
13	Report said team included an evaluation specialist	4.965	●	
27	Evaluation findings sex disaggregated at all levels	4.698	●	
15	Report indicated Conflict of Interest forms were signed	4.199	●	

As noted above, the absence of reliable data on evaluation cost and duration made it impossible for MSI to test all the hypotheses about evaluation quality that it would ideally have conducted. These two factors, along with the number of evaluation questions, are thought to collectively affect evaluation quality, as illustrated in the evaluation quality triangle on this page. MSI was, however, able to examine the third dimension of this triangle—the number of evaluation questions, or more specifically, the number of question marks in a listing of evaluation questions to see if this evaluation factor was associated with quality. The average scores for each group of evaluations, based on the numbers of evaluation questions asked to address, are in Table 77.

The results of the Chi Square test on this element showed that the number of evaluation questions and overall quality scores were not associated evaluation clusters based on numbers at a statistically significant level. The three clusters involved grouped the number of evaluation questions as 1 to 10, 11 to 20, and 21 or more. Subsequently, recognizing that USAID's How-To note on developing evaluation SOWs suggests that evaluation teams address 3 to 5 questions, MSI subdivided its initial 1 to 10 category and reran its calculations. What this showed is that when evaluations have between 1 and 5 questions, scores are marginally higher (6.33) than when they have 6 to 10 questions (6.29), but this difference was not statistically significant. While lowering the number of evaluation questions does improve overall scores, some evaluations with larger number of evaluation questions do just as well.



The fact that different numbers of evaluation questions did not affect average scores runs counter to impressions held by many evaluators and some USAID staff. This is supported by comments from the meta-evaluation's group interviews, the team leader survey, and previous studies on USAID evaluation practices.* This meta-evaluation's finding that the number of evaluation questions is not statistically associated with overall evaluation quality suggests that while a large set of evaluation questions—more than 10 or even more than 20—may theoretically be an impediment to evaluation quality, on-the-ground teams are finding ways to deal with whatever number of evaluation questions they are asked to address.

Table 77. Average Scores by Numbers of Questions

USAID Evaluation Scores 2009–12	
Evaluation Questions, (Counting Question Marks)	Average Overall Evaluation Score
1–10 questions	6.30
11–20 questions	6.58
21 or more questions	6.13

B. Overall Evaluation Scores and Evaluation Factors/Ratings Used to Construct Scores

In addition to examining associations between scores and evaluation characteristics not used in the scoring process, MSI conducted a second analysis to look at the association between scores and the quality factors used to determine those scores. This was done through a parallel set of Chi Square tests, as well as a more detailed item analysis procedure (which yielded parallel results). When Chi Square tests are used for this purpose, the resulting test values tend to be a good deal higher than when the factors examined are independent of the score. Table 78 shows the correlation between scores and items that went into creating those scores. Factors are rank ordered by degree of association.

Table 78. Quality Factors in the Overall Score Associated with the Overall Score

Factor #	Quality Rating Topic	Chi Square Value	Significance Levels		
			.05	.01	.001
35	Annex included data collection instruments	88.556	●	●	●
16	Study limitations were included	75.934	●	●	●
8	Data collection methods described	69.207	●	●	●
20	Social science methods (explicitly) were used	66.444	●	●	●
1	Executive summary mirrors report all critical elements	64.041	●	●	●
10	Data analysis method described	61.511	●	●	●
23	Findings distinct from conclusions/recommendations	53.515	●	●	●
33	SOW is included as a report annex	46.493	●	●	●
3	Project and “theory of change” described	32.438	●	●	●
32	Recommendations—clearly supported by findings	12.055	●	●	
2	Project characteristics described	10.054	●		

*See in particular, Blue & Clapp–Wincek, 2009; Hageboeck, 2009; and Frumkin & Kearney, 2010 in the meta-evaluation bibliography.

As Table 78 illustrates, nine factors are highly associated with overall evaluation scores. Two other factors that were used to construct the scoring instrument are less closely associated with overall scores. As can happen in early rounds of an index or composite scoring process, factors included in the package used to make up an index, or the evaluation quality score in this instance, may be of only tangential value. In this study, for example, the quality factor “project characteristics described” may fall in that category. However, the other less intensely associated quality factor at the bottom of this table, “recommendations are clearly supported by findings,” is less easy to dismiss. Generally speaking, the table above suggests that factors at the top of Table 78 may be predictability attributes of the most professional, highest quality evaluations USAID undertakes.

ANNEXES

Annex A. Meta-Evaluation Statement of Work

The purpose of this meta-evaluation is to provide USAID with precise information on evaluation quality aspects that are currently strengths or weaknesses; whether these aspects are improving or declining; and what opportunities exist for targeted actions that will bring overall evaluation quality into better alignment with USAID evaluation standards.

In this regard it must be noted that during the four-year period covered by this meta-evaluation, USAID has been engaged in a dramatic effort to reposition evaluation within the Agency's management cycle and establish new norms for evaluation rigor and professionalism. This evaluation initiative, arguably foreshadowed in the swearing in speech of USAID Administrator Dr. Rajiv Shah, in January 2010, has made evaluation an integral element of all aspects of the USAID programming cycle; spawned a new Evaluation Policy; and identified evaluation and improved monitoring as USAID Forward priorities.

Against this backdrop, the meta-evaluation will systematically examine a random sample of USAID evaluations from 2009-12 and gather qualitative data from USAID staff and evaluation providers to answer the following questions:

1. To what degree have quality aspects of USAID's evaluation reports, and underlying practices, changed over time?
2. At this point in time, on which evaluation quality aspects or factors do USAID's evaluation reports excel and where are they falling short?
3. What can be determined about the overall quality of USAID evaluation reports and where do the greatest opportunities for improvement lie?

To this end, MSI will, during Phase I of this meta-evaluation, develop instruments for collecting empirical data to answer these questions. The meta-evaluation will involve a systematic review of USAID evaluations produced between 2009 and 2012. The instrument employed for this review will include rating items that can clearly distinguish the difference in evaluation quality between reports, while using a multi-party rating/scoring process with a high level of inter-rater reliability standards. A detailed handbook will be needed to support the application of this instrument. The sample of evaluation reports used for the study will be representative for each of the years included in the study. Analyses of differences in evaluation reports will take place on an individual item and clustered item (topic) basis. To the extent possible, gaps between actual performance and USAID evaluation standards will be identified and reported

Data from this set of evaluations will also be compared to findings, to the extent possible, with the results of earlier reviews of USAID evaluations, most notably USAID's 1989-1991 review of 268 evaluations produced over those two years. Retrospective comparisons will also be made, to the degree possible, with one study that looked at evaluations completed in 1981-1983, and two others that examined evaluations in the 2005-2009 time frame, among others that MSI may identify as the study progresses.

To supplement and corroborate the findings from MSI's systematic document review looking at the Agency's implementation of the evaluation policy, MSI will, during Phase II of this review, gather data from USAID staff and evaluation practitioners that have been involved in post-policy evaluations. Using cost-effective approaches within the LOE established for this review, a mix of individual, group, and

focus group interviews is envisioned, for which response data will be systematically documented and analyzed using qualitative methods.

Annex B. Methods

This annex provides the reader with an in-depth understanding of the full methodology employed by MSI while designing and conducting the meta-evaluation. It is meant to elaborate on, and supplement, the methods section provided in the body of the report. In this annex you will find information on the study's conceptual framework, sampling plan, how MSI achieved interrater reliability, the data collection and analysis methods as they apply to the research questions, and the study limitations. To assist in locating specific elements of the methodology, a brief table of contents for this annex is provided below.

1. MSI's Meta-Evaluation Conceptual Framework.....	126
2. Meta-Evaluation Data Collection and Analysis Methods.....	128
3. Sampling Plan.....	132
4. Inter-Rater Reliability.....	135
5. Study Limitations.....	137

I. MSI's Meta-Evaluation Conceptual Framework

MSI's focus in this meta-evaluation is on evaluation quality; the study looks at whether the quality of USAID evaluations has changed over time. To answer questions of this nature, it is helpful to understand how evaluation quality is defined and judged. Generally speaking, those who think and write about evaluation quality tend to focus on three levels at which quality can be judged. The first, and perhaps the most aspirational level, is utilization, or, was the evaluation used and did it make a difference?³¹ The second level, which is often neglected because of the difficulties associated with making this type of judgment, is optimal design, or, did the evaluation employ the best possible design for generating the types of answers needed by evaluation clients and, by extension, was that best possible design faithfully executed?³² The third and most practical level involves the evaluation product—a report that conveys to an intended user all of the relevant information about why and how an evaluation was conducted followed by the presentation of a set of empirical findings; a set of conclusions, or interpretations, based on those findings; and an actionable list of recommendations for improving the project or program that was evaluated.

Historically, the majority of USAID's meta-evaluation work has focused on the quality of its evaluation reports, although in one instance USAID took a direct look at the utilization of its evaluations. Other development organizations that undertake periodic meta-evaluations, such as UNICEF, also tend to focus on the quality of their evaluation reports. Fitting into this tradition, the present study started with a review of the techniques used in previous USAID, and UNICEF, meta-evaluations to rate evaluation report quality. As this review showed, most meta-evaluation reviews of sets of evaluation reports use a checklist to rate a variety of evaluation elements. It also revealed that one of the first USAID meta-

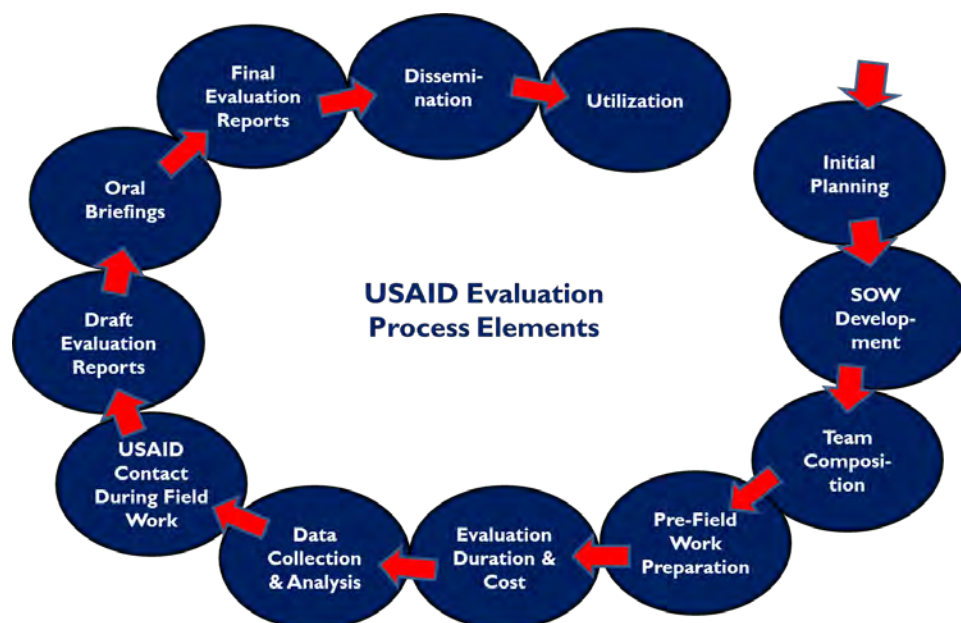
³¹ USAID undertook one evaluation review with this focus in the late 1980s, but has not reprised that effort since. Yin, Robert K.; and Carol H. Weiss. *Preliminary Study of the Utilization of AID's Evaluation Reports*. Washington DC: USAID, 1988.

³² One notable study of development program evaluations from this perspective was the Center for Global Development's study that culminated in its *When Will We Ever Learn?* report, Center for Global Development, Washington D.C. 2006. More recently, the organization *3ie* has begun to undertake systematic reviews of sets of development program evaluations that, like the CDG study focus on whether the best possible methodology was used.

evaluations (Triton, 1983) had also developed and used a single evaluation “score” to help USAID understand the messages that emerged about quality from Triton’s more detailed rating of evaluation factors.

Evaluation report quality is not a unitary concept, rather it is a collection of characteristics, many of which reflect, or are used as proxies to judge, the quality of processes carried out by evaluation teams and their clients. Some evaluation protocols for assessing evaluation quality, such as the OECD/DAC 2010 *Quality Standards for Development Evaluation*, focus on the steps in an evaluation process. For purposes of this meta-evaluation, MSI developed a visual model of that process as a data collection aid. This diagram, shown below, illustrates USAID’s evaluation process steps.

USAID Evaluation Process Wheel



Over time, a wide variety of quality checklists have been developed by evaluation theorists as well as by practitioners to assess evaluation reports from various perspectives.³³ MSI reviewed the checklists used in previous USAID meta-evaluations as a primary tool for collecting data on various aspects of evaluation quality. From MSI’s review it seems clear that while no two USAID meta-evaluations used exactly the same quality item checklist, the teams that conducted these studies were familiar with what had been done in the past, as there are a number of quality factors on which more than one meta-evaluation captured data. The results of MSI’s review of the items from prior meta-evaluation checklists are included in Annex C of this report. MSI built on this history by developing a pair of new checklists, incorporating previous elements, which were used to gather data from the sample of 340 evaluations described above.

While developing instruments and planning for the analysis of data for this meta-evaluation, MSI found it useful to think about the different evaluation actors and the roles they play throughout the evaluation process. As one moves through the Evaluation Process Wheel, some decisions are made, or heavily influenced, by the client for an evaluation while others are made primarily, if not exclusively, by

³³ Western Michigan University which offers degrees in evaluation has an Evaluation Checklist Project through which it collects and makes available a wide variety of checklists that have been used in academic courses and by project agencies to assess evaluation quality, which usually means report quality.

evaluation teams. The table below summarizes those aspects of evaluations that seem to fall primarily within the purview of the evaluation client as opposed to those that lie mainly with an evaluation team. The decisions made by these actors affect the design and implementation of an evaluation and, by extension, sections of an evaluation report that document them. Understanding which decisions are made by each evaluation actor helped the MSI team frame hypotheses about the relationships between evaluations factors, which it was then able to test during the study's data analysis period. This evaluation actor framework also helped structure recommendations from this study.

**Evaluation Quality Involves a Partnership between the Client for an
Evaluation and the Evaluation Team**

Elements an Evaluation Client Determines	Elements an Evaluation Team Provides
<ul style="list-style-type: none"> • Scope of the evaluation – single or multiple projects or programs • Timing of the evaluation – during implementation, towards the end of a project, and evaluation schedule • Management purpose – improve performance, generate lessons • Type of evaluation sought – performance, impact • Evaluation questions – number and types • Team composition – external evaluation team leader, evaluation specialist, local evaluators • Identification of deliverables, and the transmission of Agency evaluation quality standards • Duration – number of weeks or months • Evaluation budget • Evaluation quality control activities, including evaluation product reviews 	<ul style="list-style-type: none"> • Executive Summary – degree to which it accurately mirrors most critical elements of the report • Presentation of Project or Program Background – completeness from a reader's perspective • Description of the Project or Program's "Theory of Change" – development hypotheses • Presentation of the Evaluation Questions – consistency with SOW, completeness • Description of the Data Collection and Analysis Methods Used – specificity, links to questions • Description of the Study Limitation • Findings, Conclusions and Recommendations – clear distinctions among them, logical flow • Annexes – presence and completeness

2. Meta-Evaluation Data Collection and Analysis Methods

Data collection and analysis methods used for this meta-evaluation were selected based on the types of evidence needed to answer each of the three researchable questions the study sought to address, namely:

4. To what degree have quality aspects of USAID's evaluation reports, and underlying practices, changed over time?
5. At this point in time, on which evaluation quality aspects or factors do USAID's evaluation reports excel and where are they falling short?
6. What can be determined about the overall quality of USAID evaluation reports and where do the greatest opportunities for improvement lie?

These methods are summarized below on a question-by-question basis in MSI's "Getting to Answers" matrix below and are described in detail in subsections below the table. Instruments and raw data described in this methods presentation are provided separately in Annexes C, D, and E.

Meta-Evaluation “Getting to Answers” Matrix

Evaluation Question	Data Collection Methods	Data Analysis Methods
1. To what degree have quality aspects of USAID’s evaluation reports, and underlying practices, changed over time?	<ul style="list-style-type: none"> • Review of previous meta-evaluation findings • Basic characteristics coded for 340 evaluations (Instrument in Annex C) • 340 evaluations rated on 37 Quality Factors (Checklist in Annex C) • E-survey sent to 41 recent USAID evaluation team leaders (61% response rate) (Responses in Annex D) • Four small group interviews: two with USAID technical and regional office staff and two with evaluation organizations (Transcript Summaries in Annex E) 	<ul style="list-style-type: none"> • Comparative tables for current and previous meta-evaluation results • Frequency distribution for basic evaluation characteristics • Average frequency with which evaluations complied with evaluation standards on checklist overall, by year, by region, by sector and for USAID Forward evaluations • Integrated factor by factor analysis drawing on frequencies, team leader survey responses, small group interviews, and previous meta-evaluations.
2. At this point in time, on which evaluation quality aspects or factors do USAID’s evaluation reports excel and where are they falling short?	Classification of degrees of compliance into four clusters (good, fair, marginal, and weak) based on percentages of evaluations that were rated positively on 37 evaluation quality factors in checklist plus a factor for those evaluations that included ten or fewer evaluation questions	Analysis of clusters of evaluation factors to highlight where USAID excels on quality and where improvements could be made, overall, by region, by sector, and between USAID Forward and non-USAID Forward evaluations from July 2011 to December 2012
3. What can be determined about the overall quality of USAID evaluation reports and where do the greatest opportunities for improvement lie?	Construction of a ten point scale and score sheet for overall quality based on eleven quality factor checklist items. (Score Sheet in Annex C)	<ul style="list-style-type: none"> • Calculation of overall quality scores • Analysis of associations between overall quality, evaluation characteristics, and factors not used to construct the overall score • Analysis of associations between overall quality and quality factors used to construct overall scores (item analysis) • Analysis of quality factors associated with team members (team leader, evaluation specialist, and local members)

Detailed information on the data collection and analysis methods summarized above are presented on a question-by-questions basis in sections below.

Question 1: To what degree have quality aspects of USAID’s evaluation reports, and underlying practices, changed over time?

To address this question, MSI needed several types of information on the quality of evaluations both currently and in the past. Specifically, such information included a) historical, or baseline, data on USAID evaluation quality to which evaluations from MSI’s sample could be compared; b) current objective data on a yearly basis from which comparisons could be made across MSI’s study period; and c) perceptions from key stakeholders on the directionality and degree of change in evaluation quality over time.

Baseline data on USAID evaluation quality was collected through a review of previous USAID meta-evaluations the research team was able to collect. MSI’s review of these reports identified possible

evaluation quality checklist items, at the start of the meta-evaluation, and later helped identify which previous meta-evaluations included data that could be compared to findings from the current study.

Current objective data on evaluation quality factors and other characteristics were collected through the rating of the 340 USAID evaluation reports from the 2009-12 time period mentioned above in the sampling plan description. The sampling plan and sample details are presented in further detail later in section 2 below. To extract data from the evaluations in the sample, MSI used a two-part instrument. The first part was a 27 point information gathering instrument used to cull data on basic characteristics of evaluations, such as the number of evaluation questions addressed and the types of data collection methods teams used. The other was a checklist of 37 quality factors derived in part from MSI's review of past USAID meta-evaluations as mentioned above in the conceptual framework description. The data collection instrument for this aspect of the meta-evaluation, along with the coding handbook used to ensure interrater reliability can be found in Annex C.

Perceptions about whether, in what direction, and to what degree USAID evaluation quality has changed in the recent past were obtained through two different data collection methods: a team leaders' perception survey and a series of small group interviews with USAID and evaluation firms' staff. One of the pieces of information extracted from 2011-12 evaluations by MSI raters was the name of the evaluation team leader, when available. When a team leader was identified MSI then sought contact information for that individual either through Google searches, reaching out to the firms that had organized the evaluations, or other methods as necessary. Each team leader for whom contact information was located was sent an electronic Survey Monkey questionnaire. The survey was sent to 41 individuals, of which 25 responded, for a return rate of 61%. The results of this survey and the instrument are provided in Annex D. MSI also conducted two small group interviews with USAID staff. One included participants who are involved in evaluation work in regional bureaus while the other included participants from technical bureaus in USAID/Washington. In addition to interviewing USAID staff, MSI conducted two small group interviews with representatives of firms and non-governmental organizations that conduct evaluations for USAID. In all, twelve USAID staff and 25 firm representatives participated in these sessions. Each group interview used the same interview guide. Interviews were structured around a wheel depicting elements of the USAID evaluation process, shown in the conceptual framework description above. Transcripts from these sessions are provided in Annex E.

MSI's analysis of these data to answer Question 1 involved the calculation of numbers and percentages of evaluations scored by the rating team and comparisons along a number of dimensions including:

- Ratings for the current set of evaluations compared to ratings on the same or similar items in past meta-evaluations, which are summarized in small tables throughout the discussion of answers to Question 1 in the findings section of the report.
- Year-to-year comparisons for evaluations in the current study to determine how much change had occurred between 2009 and 2012, which are summarized in the findings section of this report. The summary includes MSI's calculation of the net improvement on each of the 37 items in the evaluation quality checklist and presents these checklist items rank ordered by its net increase in the percentage of evaluations rated positively on this item. This allows readers to quickly see which quality factors improve the most over the four-year study period.
- Comparisons on a sector and region basis over time, which are provided throughout the Findings section of this report. These comparisons helped to identify where factors other than time appear to affect quality ratings on particular checklist items.
- Comparisons between USAID Forward evaluations and non-USAID Forward evaluations carried out during the last 18 months of the meta-evaluation study period, which are summarized in the findings section of this report. This set of comparisons helps to clarify which evaluation quality

factors improved or failed to improve only for evaluations that received extra attention under the USAID Forward Initiative, and to identify quality factors that improved regardless of whether an evaluation was designated as a USAID Forward evaluation or not.

MSI's analysis of data collected to address Question 1 also involve the integration of data from the Team Leaders' Perception Survey and MSI's analysis of small group interview transcripts. That portion of the analysis involved reviewing each transcript and highlighting and coding participant comments. Comments on the same topics were then clustered first by topic and then by the nature of the comment offered. The frequency with which patterns were found were then documented and those frequencies along with illustrative comments were integrated into topical subsections of MSI's presentation of findings on Question 1. Once data from all three sources were drawn together, MSI was able to identify areas of convergence between sources and highlight them for readers.

Question 2: At this point in time, on which evaluation quality aspects or factors do USAID's evaluation reports excel and where are they falling short

To address this question, MSI needed information on how well USAID evaluations performed on each of the quality factors included in the rating checklist described above and provided in Annex C. While no additional data collection was required to answer this question, there was the need to employ a different data analysis process than was used in Question 1.

For Question 2, MSI's analysis of data involved the regrouping of the 37 factors from the evaluation quality checklist described above into four quality clusters based on how well USAID evaluations performed on each factor. For this analysis, MSI focused on ratings given to the set of 2012 evaluations rather than on the average rating given across four years of evaluations. The year 2012 was selected for this purpose since it most closely represents the "baseline" for future improvements in the quality of USAID evaluations. The selection of 2012 ratings as the basis for addressing Question 2 was also deemed appropriate because for 70% of the 37 evaluation quality factors MSI scored, 2012 was also the best year.

After rank ordering the 2012 evaluation factors based on the percentage of evaluations that were rated positively, or "yes," on each quality factor, MSI was able to identify four clusters indicating where USAID evaluation reports excel and where there are opportunities for improvement. These four clusters include those rated as:

- Good – 80% or more evaluations scored positively on nine evaluation quality factors in this cluster
- Fair – Between 50% and 79% of evaluations scored positively on ten evaluation quality factors
- Marginal – Between 25% and 49% of evaluations scored positively on seven evaluation quality factors
- Weak – Less than 25% of evaluations scored positively on twelve evaluation quality factors

In addition to using this clustering technique to identify where USAID evaluations excel or warrant attention in regards to specific factors, MSI also identified the sources of USAID evaluation guidance associated with each evaluation quality factor. This helped MSI spot where factors that scored weak were simply recently added quality factors therefore where a lag in adoption of new standards would not be surprising.

Question 3: What can be determined about the overall quality of USAID evaluation reports and where are the greatest opportunities for improvement?

When designing the best approach to answering this question, MSI understood that it would be best to attempt to revive the idea introduced in USAID’s 1983 meta-evaluation that if a single, composite evaluation quality “score” could be developed, and assigned to each evaluation included in a meta-evaluation, then the existence of such a score might help the meta-evaluation and USAID identify factors and characteristics that tend to be associated with relatively high evaluation quality “scores” and with relatively low “scores.”

Working with only a partial description of how quality scores were created in the 1983 meta-evaluation, MSI identified eleven items on the 37 point evaluation quality checklist described above which seemed to cover all of the major aspects of evaluation quality on which published articles and checklists on evaluation quality as well as USAID’s own guidance and meta-evaluations focus. Two of the eleven factors—program/project descriptions and the theory of change—were combined to create a ten-point scoring sheet. To ensure that any process that assigned composite quality scores to individual USAID evaluations was “fair” to all evaluations in the 2009-12 sample, MSI chose only evaluation quality factors on which USAID guidance existed as of 2008, a full year before the first evaluation in the study sample was completed. The ten evaluation quality factors MSI chose to use to create a single composite score were selected from among the full set of 37 quality factors on the checklist used by MSI raters. This list of factors was turned into a “short form” score sheet which is provided in Annex C.

As with Question 2 above, MSI did not need to gather additional data to address Question 3. The work required here involved the selection of eleven items off the 37 point check list to use to form a ten-point composite score and the calculation of scores based on those quality factors. Two of the eleven factors—program/project descriptions and the theory of change—were combined to create a ten-point scoring sheet. As it was possible for an evaluation to not receive credit for any of the quality factors, the possible scores ranged from zero to ten. The distribution of the resulting scores is described in the findings section under Question 3 of this report. In addition to calculating these scores, MSI prepared cross-tabulations and calculated chi square values to determine the degree to which scores were associated with specific quality factors or other evaluation characteristics on which MSI raters had already gathered data. With scores assigned to individual evaluations, MSI was then able to determine the average scores for clusters of evaluations, such as USAID Forward and non-USAID Forward evaluations, and run t-tests to determine whether differences between group averages were statistically significant. Through these processes, MSI was able to identify a number of factors which are associated with relatively high, and relatively low, composite evaluation quality scores.

As a final element of its overall analysis, MSI was able to take the findings from meta-evaluation questions and relate them to the steps in the Evaluation Process Wheel and the evaluation roles outlined in the conceptual framework description above in order to best frame the recommendations from this study that are provided in the final section of this report.

3. Sampling Plan

With the conceptual framework in place, MSI developed a sampling plan to support detailed year-by-year comparisons of the quality of USAID evaluation reports over the period 2009-2012. The following section provides information on the universe of evaluations for that period as well as MSI’s sampling plan and firewalls established to guard against any potential conflicts of interest or security concerns.

Evaluation Universe

On February 7, 2013, USAID/PPL/LER provided MSI with information on 624 evaluation reports produced between 2009 and 2012 that were coded as “final evaluations” or “special evaluations” in the USAID Development Experience Clearinghouse (DEC). This set excluded any evaluations that were specified in the DEC as foreign language. Functionally, this set represented the universe of evaluations for the study. A spreadsheet covering these reports was prepared by the DEC and transferred to MSI via USAID. The reports were delivered in two Excel files: one each for final and special evaluations. MSI took these two files and combined them, merging final and special evaluations into four year-by-year spreadsheets. In March 2013, USAID provided information on an additional eight evaluations to be included in the study; these were USAID Forward evaluations that fell within the appropriate time frame but were not originally included in the evaluations received from the DEC. The information on these eight evaluations was incorporated into the relevant spreadsheets which increased the study’s universe to 632 evaluations.

This study’s universe represented 128 out of 187 evaluations completed in 2011 or 2012 that the Agency had designated under its USAID Forward target for producing high quality evaluations. The remaining USAID Forward evaluations were either completed in 2013 or written in a foreign language. MSI’s spreadsheet of evaluations specifically identified those that have a USAID Forward designation.

While the majority of the evaluations were available through the DEC, there were 24 evaluations designated as “restricted documents” as they contained Sensitive but Unclassified (SBU) information. These restricted documents are not publicly available and therefore are not accessible through the DEC. USAID provided MSI with these documents on an “as-needed” basis in PDF format.

Evaluation Sample

From this universe, MSI drew separate samples for each year, using an 85% confidence level and +/- 5% confidence interval to determine sample sizes per year. These parameters were selected with an eye to being both indicative of our certainty with respect to study findings and mindful of the budget available for this task. Additionally, by setting these parameters on an annual basis, MSI was also able to reach a 99% confidence level and +/- 5% confidence interval for the full sample across all four years.

Based on past experience with USAID meta-evaluations, MSI anticipated that roughly 10% of the documents received from the DEC would turn out to be something other than evaluations. The original sample size table, by year, show below, takes this 10% non-evaluations in the DEC universe into account.

Year	Number of DEC Documents Coded as Evaluations (Final or Special)	Number of Anticipated Evaluations Assuming 10% DEC Coding Error	Sample at 85%, +/- 5	Statistical Characteristics of Samples Drawn Separately for Each Year of the Meta Evaluation
2009	124	112	73	For each sample year: Confidence Level: 85% Margin of Error: +/- 5%
2010	153	138	84	
2011	178	160	91	
2012	177	159	91	
Combined	632	569	337	Confidence Level: > 99% Margin of Error: +/- 5%

To implement this sampling plan, MSI used a randomization function within Excel to assign a randomized number to every evaluation. The function used is the “Rand ()” function which assigns a random

number between 0 – 1 (e.g., .3854924783) to each cell indicated. Once random numbers were assigned, MSI reorganized all of the information in the randomized numbered rows by sorting the random number column from lowest to highest. This produced a spreadsheet of randomly organized evaluations and associated data. The process was repeated for each of the four spreadsheets corresponding to the four years.

With the four spreadsheets organized in the randomized fashion described, MSI was then able to select the evaluations for our study by starting at the top of the spreadsheet and moving down the sheet sequentially (and therefore in a random selection order) until the quota for each year's sample was filled (73 for 2009, 84 for 2010, 91 for 2011, and 89 for 2012). All evaluations not included in the MSI set were saved on file in their randomized state. In the case where some in the original set needed to be excluded, i.e., where non-evaluations were discovered by the coders, these were replaced by new evaluations in the same randomized sequential order.

After all 337 were scored, MSI found that fewer “non-evaluations” were present than originally expected. The actual percentage of non-evaluations was closer to seven percent than the predicted 10% as shown in the following table.

Year	Number of DEC Documents Coded as Evaluations (Final or Special)	Number of DEC Documents Verified as Evaluations	Percentage of DEC Documents in Sample Validated as Evaluations
2009	124	112	90%
2010	153	142	93%
2011	178	154	87%
2012	177	165	93%
Total	632	573	

For this reason, MSI adjusted the sample size for each year proportionally to ensure maintaining an 85% confidence level and +/- 5% confidence interval per year and a 99% confidence level and +/- 5% confidence interval for the full sample across all four years. The following table provides the final sample sizes by year in relation to the actual number of verified evaluations in the DEC universe.

Year	Number of DEC Documents Verified as Evaluations	Number of Verified Evaluations Coded in the Meta-Evaluation	Statistical Characteristics of Samples Drawn Separately for Each Year of the Meta Evaluation
2009	112	73	For each sample year: Confidence Level: 85% Margin of Error: +/- 5%
2010	142	85	
2011	154	89	
2012	165	93	
Combined	573	340	Confidence Level: > 99% Margin of Error: +/- 5%

Firewalls

Once year-by-year samples were selected, MSI implemented three distinct firewalls. The first was to identify all evaluations within the sample that were flagged by USAID as “restricted documents” and therefore contained Sensitive but Unclassified (SBU) information. All SBU evaluations were coded exclusively by members of the study team with active USG secret level security clearances or higher. A total of 13 such evaluations were included in the original sample. In only one instance was an additional SBU evaluation incorporated into the sample to replace a non-evaluation, making the total number of restricted documents 14.

The same process was conducted for the second firewall, which was set in place to prevent an MSI employee from coding an evaluation report authored by, or with contributions from, MSI, which would be considered a conflict of interest.

The third firewall was set in place in response to one of the coders who was an external consultant, but had previously worked on an MSI contract in South Sudan to manage evaluations taking place there. This firewall prevented the coder from reviewing any evaluations of projects in South Sudan. To address all MSI/South Sudan evaluations, MSI had an additional external consultant with no connection to South Sudan that was assigned all of these evaluations. Unfortunately, this consultant was not able to complete the study, so these evaluations were assigned to an MSI employee that works in the Human Resources department and was considered to be sufficiently removed from the evaluation practice at MSI, and in particular the evaluations conducted in South Sudan, to the point that the integrity of this firewall would be maintained.

There were two evaluations found to be both MSI evaluations and SBU evaluations. Since all team members with active clearances were MSI staff, and therefore could not code an evaluation of this type, it was decided that a USAID representative would be trained and would score these two evaluations in the same manner as the rest of the team. This same USAID representative also coded a third evaluation when a situation was encountered where no MSI employee or consultant on the coding team could code the evaluation without some form of bias.

4. Interrater Reliability

Having established the conceptual framework and designing a sampling plan, MSI then turned to what it recognized early on as one of the most critical elements of a meta-evaluation of this scope, namely interrater reliability (IRR) among coders. With nearly 350 evaluations to code within a two to three month window, MSI originally anticipated having a team of eight coders working half-time or more for eight weeks. The reality ended up being a team of ten coders working for closer to 12 weeks. To promote IRR, MSI utilized a collaborative process of group discussions, group scoring exercises, and periodic IRR Tests. This process allowed the evaluation team to identify challenges and come to consensus on checklist items.

The instrument used for evaluation coding included a checklist and a corresponding handbook detailing how coders should interpret each item on the checklist. Once MSI created an initial draft of the checklist and handbook it was shared with, and approved by, USAID. The understanding with USAID was that while the language might be altered, the general purpose of the checklist items and the data it would represent would not change. While MSI put a great deal of effort into reducing subjectivity in checklist items, MSI and USAID recognized that subjective elements could not be eliminated completely.

To maximize the potential for IRR, once USAID approved the structure and content of the instrument, MSI engaged the full team of coders in the refinement of the instrument. Coders received the checklist and corresponding handbook as well as the 2011 USAID Evaluation Policy and current ADS 203 in

advance of the first team meeting. The initial team meeting was meant to introduce the instrument and provide instructions to the coders on how each checklist item should be interpreted and how appropriate responses should be determined in various circumstances. This first meeting sparked longer conversations than originally anticipated and was turned into a two-part meeting to ensure that each issue was sufficiently addressed. Throughout the meetings, coders identified a number of items that required refinement and clarification.

Following the second meeting, the coders were collectively assigned a single evaluation to score using the modified checklist; this evaluation was a USAID evaluation, but not one included in our sample for the study. The evaluation manager analyzed results from this pretest and then brought the coding team together to discuss coding differences and any issues with the application of the instrument. This discussion resulted in the revision of a number of additional items; several questions were identified as too subjective and therefore removed; and compound questions were split into two questions. MSI shared the changes with USAID as the instrument evolved.

The team demonstrated increased cohesiveness of thought in a second full-team pretest; the team reconvened to discuss the second pretest and the instrument was further refined. At this point the Evaluation Team Leader divided the team into smaller groups for a third pretest.³⁴ The purpose of this approach was to increase confidence and independence of coders while continuing to encourage discussion on how items should be coded in specific circumstances. It was during this third pretest round that coders began reading evaluations from the study's sample. Either the Evaluation Team Leader or Evaluation Team Manager also read these reports to ensure the accuracy of data, answer any questions, and gather and share lessons learned or best practices with the rest of the coders. Some minor adjustments were made to the instrument.

For the fourth pretest, coders were placed in new teams. The mixing of teams was intended to increase the familiarity and comfort level among coders so that once on their own they would continue to engage with their colleagues to identify and work through issues, thereby maintaining and increasing IRR through a group-thought mentality. At this stage three coders were prepared to work independently. The remaining four were again paired up and provided a fifth pretest. The Evaluation Team Manager was present during the discussion of the fourth and fifth pretests, overseeing the process and identifying information to share to the rest of the team. Minimal adjustments were made to the instrument during these two stages, and updates were emailed to all coders.

With all seven coders having successfully completed their pretests, they began coding. Coders were instructed to talk amongst themselves to work through any questionable items on specific issues and to share their thoughts via emails to ensure the group continued to think and troubleshoot issues in the same manner. Should an issue arise in which a group of coders could not come to consensus, the issue was raised with the Evaluation Team Manager who instructed the entire team on how to address such issues in the future. If the Evaluation Team Manager was unable to definitively resolve an issue, the issue was then brought to the attention of the Evaluation Team Leader who made the final decision, at which point the decision, and the reasoning behind the decision, were shared with the coding team. In most instances, coders were directed to the exact language used in the Evaluation Policy or the ADS 203 that addressed the issue.

Shortly after coders began working individually, a second coder was pulled away from the study and onto another project. It was at this point, already four weeks into the process, that the Evaluation Team Manager identified two potential candidates to join the team. The two candidates read the ADS 203, Evaluation Policy, and the Instrument, and then discussed the study and the Instrument with the

³⁴ One of the coders unexpectedly had to leave the team at this point so the coders were divided into two teams of two and one team of three.

Evaluation Team Manager. The candidates were then assigned a pretest used by the other coders so that their responses could be compared to the rest of the team. Each candidate tested well, asked appropriate questions, and moved on to a second pretest.

To optimize IRR and incorporate the new coders into the team, MSI combined the new coder's second pretest with an IRR test for the original coders. This also coincided with the point at which 25% of the evaluations had been coded. All eight coders were assigned one single evaluation from the study sample. The responses were analyzed to identify the number of deviations each coder had from the appropriate response.³⁵ It was decided that coders who deviated from the group 10% or less of the time could continue coding, while coders who deviated more than 10% of the time were either put on probation or removed from the study.³⁶ Following the first IRR test, both new coders were welcomed on to the team, six existing coders continued coding, one coder was put on probation, and one coder was let go.

The second part of the analysis from the IRR test was to identify specific checklist items on which there was the most disagreement. Items with one or two deviations were considered acceptable, while items with three or more deviations were considered problematic. If such questions existed and were discussed in greater detail as a team. For all II items either the wording of the item was changed or the description of how to interpret the item in the handbook was changed; changes were only made on consensus. Following changes, all coders were asked to revisit previous evaluations to confirm the correct codings were made and to change their responses as necessary for these items.

To replace the coder who was let go after the first IRR test, and to increase the pace of the study, the evaluation team manager identified three new coder candidates. These three candidates were brought on line in the same manner as the previous candidates with the provision of the Evaluation Policy, the ADS 203, and the instrument, followed by a meeting with the evaluation team manager, and two pretests, one of which was the first IRR test. While all three candidates made it to the second pretest, only one candidate had less than 10% deviations and was brought on to the team. The same process also took place with a USAID representative who was asked to code two evaluations, which no coder on the team could read due to firewalls put in place to prevent bias and protect classified material. The USAID representative scored well on the pretests and moved on to become a coder.

With 65% of the evaluations completed, the evaluation team leader decided to conduct a second IRR test and to attempt to bring on two more coders. A single evaluation was assigned to the full team and the responses were analyzed in the same manner as before. In addition to the seven main coders, two previous coders were using this IRR test to determine whether they could rejoin the team; one had been pulled away from the study for several weeks while the other had been on probation. Based on the analysis of results, all nine coders were found to have less than 10% deviations and were cleared to continue coding. Additionally, one of the two candidates achieved less than 10% deviations and was welcomed on to the team. The IRR test also identified seven problematic items which were addressed through the rewording of items and clarifications provided in the handbook; one item was removed from the checklist at this point. All coders were asked to revisit previous evaluations at this time to ensure accurate responses.

Following the second IRR test, the evaluation team leader and evaluation team manager convened another full team meeting to review initial data from the coding process. Coders were asked to look for any anomalies or surprising pieces of data. Coders identified a few questions which were further

³⁵ The appropriate response was typically determined by the majority, but in select cases it was determined that the minority were in fact closer to the correct response.

³⁶ Probation, in this sense, means that the coder had to speak with the Evaluation Team Leader about how they can better bring themselves in line with the group; meet with the Evaluation Team Manager to discuss each deviation in detail; shadow another coder for at least one evaluation; and wait until the next IRR test.

discussed and led to rewording in the checklist and further clarification in the handbook. Coders were asked to revisit select evaluations they previously coded to ensure accurate data was being collected.

5. Study Limitations

Study limitations for this meta-evaluation include a variety of data situations which may compromise the accuracy of the study's findings. Limitations are discussed below by source.

Development Experience Clearinghouse

- USAID requires that all evaluations be submitted to the DEC, and the USAID Evaluation Policy modifies this requirement to state that submission must take place within three months of completion. However, there is, as of yet, no mechanism for determining adherence to this requirement. It is thus possible that additional evaluations from the study time period exist which were not included in the population sampled.
- USAID policy does not stipulate who is responsible for uploading evaluations or the process for doing so. This has led to documented inconsistencies in the data available on the DEC, including: duplication of evaluations on the DEC; non-evaluations being uploaded and labeled as evaluations; and omitted or inaccurate descriptive data for evaluations. During the rating process, raters were asked to indicate whether descriptive information from the DEC was accurate for five descriptive elements; the team identified that for these five elements, there were inaccuracies between 3% and 10% of the time, depending on the element. While many inaccuracies were identified, other unfound inaccuracies may have affected some of the findings for this study, meaning that there is potential for data being inaccurate due to misinformation provided by the DEC.

Checklists and Inter-Rater Reliability

- While MSI is confident that subjective items in the checklist have been removed and that there is strong inter-rater reliability among coders, we recognize that both of these issues are inherent problems with a study of this kind and may have affected some of the data and findings.

Cost and Duration Data

- The relationship between the number of evaluation questions, the budget for the evaluation, and the duration of the evaluation has long been cited as one of the most critical factors in measuring the quality of evaluations. In the 1983 Triton study, data on all three factors was available and used in data analysis, and the 1987-1989 meta-evaluation reported that cost data were available for 45% of the 287 evaluations examined for those years. As this meta-evaluation was asked to be as comprehensive as possible, and was meant to compare to historical data, MSI hoped to be able to work with cost and duration data, even if just for a subset of the sample. Though USAID put forth a great effort to locate such data, and ultimately was able to find some, the data available were unreliable and therefore not able to be used in this study. Without information on cost and duration, MSI was unable to test for the degree of association between these three factors or other pairs of evaluation quality variables, which limited the findings.

Team Leader Survey

- Of the evaluations in our sample from 2011-2012, 60% (112/184) did not identify the evaluation team leader. As a result, MSI was unable to identify a person to include in the survey sample.

- Of the 41 Team Leaders to whom the survey was sent, three respondents started but did not complete the survey with one indicating that it was the result of poor internet while traveling; it can be assumed that remaining two incomplete respondents had similar complications.
- 25 respondents is a fairly small number to use to characterize all recent USAID evaluation team leaders, regardless of the size of that population. For this reason MSI's presentation of data from this survey was couched in language that was designed to encourage the reader to listen to the voices of those team leaders without trying to generalize what was heard to a larger universe.

Group Interviews

- Small group interviews carried out to elicit information are subject to limitations with respect to their generalizability. Several issues of this sort are noted below.
 - Of the Technical Bureau Representatives invited to participate, seven people out of fourteen attended and only represented three of the bureaus and offices invited, representing less of the sectors than desired, but still providing valuable insights;
 - Of the Regional Bureau Representatives invited to participate, five people out of nine attended and only represented five of the bureaus and offices invited, representing less of the regions than desired, but still providing valuable insights; and
 - Of the 24 firms invited, four declined participation, three never responded, and one did not show up.

Mission Staff Input

- To provide as complete a picture of as possible of evaluation quality in USAID, MSI hoped to be able to obtain inputs for the meta-evaluation from field staff. In the end that was not possible due to limitations within USAID on the number of surveys and other data collections it felt Missions could be burdened with. As a result, the meta-evaluation lacks in-depth client data on the evaluator-determined elements of evaluations that it was able to obtain for client-determined evaluation factors from small group interviews and its team leader survey.

Annex C. Meta-Evaluation Rating Instruments

This annex provides the reader with the instruments used for collecting quantitative data from the evaluations included in the meta-evaluation's sample. Included here are not only the checklist for rating evaluations and the basic evaluation characteristics collection instrument, but also the instrument used to develop those instruments and the handbook used by MSI to ensure inter-rater reliability. To facilitate ease of finding each instrument within this annex, a small table of contents for this annex is provided below

1. Historical Analysis of Coverage and Sources of Questions from Previous USAID Meta Evaluations.....140
2. Basic Evaluation Characteristics Description Instrument147
3. Evaluation Rating Checklist Instrument.....153
4. Short Form: Evaluation Ten Point Score Elements.....156
5. Basic Evaluation Characteristics Description Instrument Handbook.....157
6. Evaluation Rating Checklist Instrument Handbook.....164

I. Historical Analysis of Coverage and Sources of Questions from Previous USAID Meta Evaluations.

A. Basic Characteristics	Triton (1982)	Triton (1983)	Development Associates (1987-1988)	MSI (1989-1990)	Clapp-Wincek & Blue (1998-1999)	Bollen (2005)	MSI (2005-2008)	MSI (2008 remaining months)	Kumar and Eriksson (2009)
1. What kind of document is it? (Evaluation, Assessment, Audit, etc.)				•	•		•	•	
2. Year Published	•	•	•	•	•		•	•	
3. Month the Report was Published (enter the month, e.g., May)									
4. Document Title				•					
5. Authorizing Organization									
6. Sponsoring Organization	•			•					
7. Geographic Descriptors	•	•	•	•			•	•	•
8. Primary Subject				•				•	
9. USAID Evaluation Activity Manger									
10. Report Length									
a. Executive Summary									
b. Report, including Executive Summary, excluding annexes									

META-EVALUATION OF QUALITY AND COVERAGE OF USAID EVALUATIONS 2009–12

A. Basic Characteristics	Triton (1982)	Triton (1983)	Development Associates (1987-1988)	MSI (1989-1990)	Clapp-Wincek & Blue (1998-1999)	Bollen (2005)	MSI (2005-2008)	MSI (2008 remaining months)	Kumar and Eriksson (2009)
11. Evaluation Type (Performance, Impact, Hybrid, etc.)								●	
12. Timing (During Implementation, Towards end, Ex-Post, etc)	●		●	●	●		●	●	●
13. Scope (Single project, multi-project, program, etc.)			●	●			●	●	
14. Evaluation Purpose – policy list - only if explicitly stated (Learning, Accountability, or Both)				●	●				
15. Specific Evaluation Purpose Included in Report from the list provided				●	●		●	●	
16. What was the evaluation asked to address? (Questions, Issues, other)									
17. Number of evaluation questions									
a) Are the questions numbered? Yes or no?									
b) Highest number assigned, (e.g., 5) even if there were a number of sub-questions									
c) Count of all question marks, including in sub-questions									
d) Considering all questions, including when you split up compound questions (two questions with an “and,” but only one question mark?)									
18. Evaluation Design/Approach to Causality/Attribution									

META-EVALUATION OF QUALITY AND COVERAGE OF USAID EVALUATIONS 2009–12

A. Basic Characteristics	Triton (1982)	Triton (1983)	Development Associates (1987-1988)	MSI (1989-1990)	Clapp-Wincek & Blue (1998-1999)	Bollen (2005)	MSI (2005-2008)	MSI (2008 remaining months)	Kumar and Eriksson (2009)
Included									
19. Specific Design for Examining Causality/Attribution the Team Used								•	
20. & 21. Data Collection methods (documents, interviews, surveys, other options cited and/or actually used)		•	•	•		•	•	•	•
22. Data Analysis methods planned				•					
22. Data Analysis methods used				•					
24. Did the evaluation report state that a participatory approach or method was used?								•	
25. Participatory – who participated (check all that apply)								•	
26. Participatory – phase of evaluation (check all that apply)								•	
27. Number of Recommendations									

B. Evaluation Rating Checklist	Triton (1982)	Triton (1983)	Development Associates (1987-1988)	MSI (1989-1990)	Clapp-Wincek & Blue (1998-1999)	MSI (2008-2010)	Kumar and Eriksson (2009)	MSI Aid to Trade (2002-2010)	MSI CEC Courses	Current USAID
Executive Summary										
1. Does the Executive Summary accurately reflect the most critical elements of the report?						•		•	•	•
Program/Project Background										
2. Are the basic characteristics of the program, project or activity described (title, dates, funding organization, budget, implementing organization,				•		•	•	•	•	•

META-EVALUATION OF QUALITY AND COVERAGE OF USAID EVALUATIONS 2009–12

B. Evaluation Rating Checklist	Triton (1982)	Triton (1983)	Development Associates (1987-1988)	MSI (1989-1990)	Clapp-Wincek & Blue (1998-1999)	MSI (2008-2010)	Kumar and Eriksson (2009)	MSI Aid to Trade (2002-2010)	MSI CEC Courses	Current USAID
location/map, target group)?										
3. Is the program or project's "theory of change" described (intended results (in particular the project purpose); development hypotheses; assumptions)						●		●	●	●
Evaluation Purpose										
4. Does the evaluation purpose identify the management reason(s) for undertaking the evaluation?				●		●	●	●	●	●
Evaluation Questions										
5. Are the evaluation questions clearly related to evaluation purpose?										
6. Are the evaluation questions in the report identical to the evaluation questions in the evaluation SOW?						●	●	●	●	●
7. If the questions in the body of the report and those found in the SOW differ, does the report (or annexes) state that there was written approval for changes in the evaluation questions?										
Methodology										
8. Does the report (or methods annex) describe <u>specific</u> data collection methods the team used?				●		●	●	●	●	●
9. Are the data collection methods presented (in the report or methods annex) in a manner that makes it clear which specific										

META-EVALUATION OF QUALITY AND COVERAGE OF USAID EVALUATIONS 2009–12

B. Evaluation Rating Checklist	Triton (1982)	Triton (1983)	Development Associates (1987-1988)	MSI (1989-1990)	Clapp-Wincek & Blue (1998-1999)	MSI (2008-2010)	Kumar and Eriksson (2009)	MSI Aid to Trade (2002-2010)	MSI CEC Courses	Current USAID
methods are used to address <u>each</u> evaluation question? (e.g., matrix of questions by methods)										
10. Does the report (or methods annex) describe <u>specific</u> data analysis methods the team used? (frequency distributions, cross-tabulations; correlation; reanalysis of secondary data)										
11. Are the data analysis methods presented (in the report or methods annex) in a manner that makes it clear how they are associated with the evaluation questions or specific data collection methods?										
Team Composition										
12. Did the report (or methods annex) indicate that the evaluation team leader was external to USAID?										
13. Did the report (or methods annex) identify at least one evaluation specialist on the team?					•					
14. Did the report (or methods annex) identify local evaluation team members?	•	•	•	•	•					
15. Did the report indicate that team members had signed Conflict of Interest forms or letters? (check if the report says this or the COI										

META-EVALUATION OF QUALITY AND COVERAGE OF USAID EVALUATIONS 2009–12

B. Evaluation Rating Checklist	Triton (1982)	Triton (1983)	Development Associates (1987-1988)	MSI (1989-1990)	Clapp-Wincek & Blue (1998-1999)	MSI (2008-2010)	Kumar and Eriksson (2009)	MSI Aid to Trade (2002-2010)	MSI CEC Courses	Current USAID
<i>forms are included in an annex)</i>										
Study Limitations										
16. Does the report include a description of study limitations (lack of baseline data; selection bias as to sites, interviewees, comparison groups; seasonal unavailability of key informants)?						•	•	•	•	•
Responsiveness to Evaluation Questions										
17. Is the evaluation report structured to present findings in relation to evaluation questions, as opposed to presenting information in relation to program/project objectives or in some other format?						•		•	•	
18. Are <u>all</u> of the evaluation questions, including sub-questions, answered primarily in the body of the report (as opposed to in an annex)										
19. If any questions were not answered, did the report provide a reason why?										
Findings										
20. Did the findings presented as the basis for answering evaluation questions appear to be drawn from social science data collection and analysis methods the team described in its study methodology										

META-EVALUATION OF QUALITY AND COVERAGE OF USAID EVALUATIONS 2009–12

B. Evaluation Rating Checklist	Triton (1982)	Triton (1983)	Development Associates (1987-1988)	MSI (1989-1990)	Clapp-Wincek & Blue (1998-1999)	MSI (2008-2010)	Kumar and Eriksson (2009)	MSI Aid to Trade (2002-2010)	MSI CEC Courses	Current USAID
(including secondary data it assembled or reanalyzed)?										
22. In the presentation of findings, did the team draw on data from the range of methods they used rather than answer using data from primarily one method?										
23. Are findings clearly distinguished from conclusions and recommendations in the report, at least by the use of language that signals transitions (“the evaluation found that.....”, “the team concluded that”)?		•	•	•		•		•	•	•
24. Are quantitative findings reported precisely, i.e., as specific numbers or percentages rather than general statements like “some”, “many”, or “most”?						•		•	•	•
25. Does the report present findings about unplanned/unanticipated results?				•						
26. Does the report discuss alternative possible causes of results/outcomes it documents?				•			•		•	•
27. Are evaluation findings disaggregated by sex at all levels (activity, outputs, outcomes) when data are person-focused?			•	•						
28. Does the report explain whether										

META-EVALUATION OF QUALITY AND COVERAGE OF USAID EVALUATIONS 2009–12

B. Evaluation Rating Checklist	Triton (1982)	Triton (1983)	Development Associates (1987-1988)	MSI (1989-1990)	Clapp-Wincek & Blue (1998-1999)	MSI (2008-2010)	Kumar and Eriksson (2009)	MSI Aid to Trade (2002-2010)	MSI CEC Courses	Current USAID
access/ participation and/or outcomes/benefits were different for men and women when data are person-focused?										
Recommendations										
29. Is the report's presentation of recommendations limited to recommendations? (free from repetition of information already presented or new findings not previously revealed)				•		•		•	•	
30. Do evaluation recommendations meet USAID policy expectations with respect to being specific? (states clearly what is to be done, and possibly how?)										
31. Do evaluation recommendations meet USAID policy expectations with respect to being directed to a specific party? (identifies who should do it)						•		•	•	•
32. Are all the recommendations supported by the findings and conclusions presented? (Can a reader can follow a transparent path from findings to conclusions to recommendations?)						•	•	•	•	•
Annexes										
33. Is the evaluation SOW included as an annex to the	•		•	•	•				•	•

META-EVALUATION OF QUALITY AND COVERAGE OF USAID EVALUATIONS 2009–12

B. Evaluation Rating Checklist	Triton (1982)	Triton (1983)	Development Associates (1987-1988)	MSI (1989-1990)	Clapp-Wincek & Blue (1998-1999)	MSI (2008-2010)	Kumar and Eriksson (2009)	MSI Aid to Trade (2002-2010)	MSI CEC Courses	Current USAID
evaluation report?										
34. Are sources of information that the evaluators used listed in annexes?										
35. Are data collection instruments provided as evaluation report annexes?										
36. Is there a matching instrument for <u>each</u> and <u>every</u> data collection method the team reported that they used? ³⁷							●		●	●
37. Were any “Statements of Differences” included as evaluation annexes (prepared by team members, the Mission, the Implementing Partner, or other stakeholder)?										
Evaluation Data Warehousing										
38. Does the evaluation report explain how the evaluation data will be transferred to USAID (survey data, focus group transcripts)?										
SOW Leading Indicator of Evaluation Quality (if SOW is a report annex)										
39. Does the evaluation SOW include a copy or the equivalent of Appendix I of USAID’s evaluation policy?										

³⁷ Though removed from other checklists and analyses in the report due to inconsistency and unreliability of the data from this element, it is left here to identify why MSI attempted to collect data on this element in the first place.

2. Basic Evaluation Characteristics Description Instrument

Basic Evaluation Characteristics	Answer in this Column Y/N or text
1. What kind of document is it? (Select only one option)	
• Evaluation	
• Audit (IG or GAO)	
• Assessment	
• Meta-analysis	
• Meta-evaluation	
• Evaluation guidance	
• Other <i>Please insert exact language from there report here.)</i>	
• Unable to determine	
If this document is not an evaluation, STOP HERE.	
2. Year Published (read spreadsheet and confirm, if correct enter Yes to the right, if No, enter correct answer directly below)	
3. Month the Report was Published (enter the month, e.g., May	
4. Document Title (answer as above)	
5. Authorizing Organization (answer as above)	
6. Sponsoring Organization (answer as above)	
7. Geographic Descriptors (answer as above)	
8. Primary Subject (answer as above)	
9. USAID Evaluation Activity Manger (enter or paste name below)	
10. Report Length	
c. Executive Summary <u>alone</u> (pages)	
d. Report, including Executive Summary, excluding annexes (pages = final page number for body of the report)	
11. Evaluation Type (choose only one)	
• Performance	
• Impact	
• Both (hybrid)	
• Unable to determine	
12. Timing (choose only one)	
• During Implementation	
• Towards End of Program/Project	
• Continuous (parallel Impact Evaluation)	

Basic Evaluation Characteristics	Answer in this Column Y/N or text
• Ex-Post	
• Unable to determine	
13. Scope (choose only one)	
• Single Project or activity (one country)	
• Program-level (one country) – explicitly examines all elements under a USAID Development Objective (DO), e.g., “economic growth improved”, “food security increased”	
• Sector-wide (one country) – e.g., all agriculture, all health projects/activities	
• Other Multiple Projects (one country) evaluation, e.g., several activities in one district, or several activities focused on youth employment	
• Single project (multiple countries) e.g., approach to sexual violence in schools in Ghana and Malawi	
• Multiple projects (multiple countries), e.g., worldwide review of Mission funded trade projects	
• Regional program or project (funded by a regional office or bureau); e.g., Mekong River cooperation project involving multiple countries	
• Global program or project (funded by USAID/W), e.g., worldwide assistance to missions on gender assessments	
• Other scope (<i>explain or paste in description below</i>)	
• Unable to determine	
14. Evaluation Purpose – policy list - only if explicitly stated (choose one) one	
• Learning	
• Accountability	
• Both	
• Neither one was explicitly stated	
15. Specific Evaluation Purpose Included in Report	
Data capture: Insert the exact Evaluation Purpose language from the report at right	
Check <u>all that apply</u> below regarding the Evaluation Purpose, i.e., management reason(s) for undertaking the evaluation	
a) Improve the implementation/performance of an existing program, project, or activity	
b) Decide whether to continue or terminate an existing project or activity	
c) Facilitate the design of a follow on project or activity	
d) Provide input/lessons for the design of a future strategy, program, or project that is not a direct follow-on (i.e., not Phase II) of the one this evaluation addressed.	
e) Required by policy, i.e., performance evaluations of large projects or impact evaluations of innovative interventions or pilot projects	
f) USAID Forward commitment (Mission commitment to produce a specific number of USAID Forward evaluations)	
g) Other (<i>explain or paste purpose statement below</i>)	

Basic Evaluation Characteristics	Answer in this Column Y/N or text	
h) Unable to determine		
16. What was the evaluation asked to address?		
a) Questions, Issues, Other (for “other” <i>explain or paste in description below</i>), or you can indicate that the evaluation was not asked to address anything in particular		
Other:		
17. Number of evaluation questions		
e) Are the questions numbered? Yes or no?		
f) Highest number assigned, (e.g., 5) even if there were a number of sub-questions		
g) Count of all question marks, including in sub-questions		
h) Considering all questions, including when you split up compound questions (<i>two questions with an “and,” but only one question mark?</i>)		
18. Evaluation Design/Approach to Causality/Attribution Included		
<ul style="list-style-type: none"> Did the list of evaluation questions include questions about causality/attribution? If no, skip Question 18 below. 		
19. Specific Design for Examining Causality/Attribution the Team Used	Y/ N or N/A	
a) The evaluation report says it used an <u>experimental design</u> or provided equivalent words (control group, randomized assignment, randomized controlled trial). If yes, enter “yes” and provide the page number.		If yes, provide page number
b) The evaluation report says it used a <u>quasi-experimental design</u> or provided equivalent words (comparison group, regression discontinuity; matching design; propensity score matching, interrupted time series). If yes, enter “yes” and provide the page number.		If yes, provide page number
c) The evaluation report says it used a specific <u>non-experimental approach</u> for examining causality or attribution (outcome mapping; identification & elimination of alternative possible causes (<i>modus operandi</i>); contribution analysis, case study). If yes, enter “yes” and provide the page number.		If yes, provide page number
d) While there were questions about causality/attribution in the list, no overall design for answering these questions was presented		
Data Collection methods (check all that apply)	20. Methods section said planned to use the method to collect data	21. Findings presentation explicitly references data from this method
a) Cull data from non-project document review/secondary data sources		
b) Cull facts from project-related documents/data sources		
c) Structured observation		
d) Unstructured observations		
e) Key Informant interviews		
f) Individual interviews		
g) Survey		
h) Group interviews		

META-EVALUATION OF QUALITY AND COVERAGE OF USAID EVALUATIONS 2009–12

Basic Evaluation Characteristics	Answer in this Column Y/N or text	
i) Focus group		
j) Community interview/town hall meeting		
k) Instruments – weight, height, pH		
l) Other data collection method (describe or paste in below)		
m) Unable to determine		
Data Analysis methods (check all that apply)	22. Methods section said the team planned to use the method to analyze data	23. Visible use, or explicit reference to results from this method
a) Descriptive statistics (frequency, percent, ratio, cross-tabulations)		
b) Inferential statistics (regression, correlation, t-test, chi-square)		
c) Content or pattern analysis (describes patterns in qualitative responses)		
d) Other data analysis method (describe or paste in below)		
e) Unable to determine		
24. Did the evaluation report state that a participatory approach or method was used? If yes, indicate who participated (beyond contributing data) and at what stage of the evaluation in questions 25 and 26 below. If not, skip questions 25 and 26.		
25. Participatory – who participated (check all that apply)		
a) USAID staff		
b) Contractor/grantee partner staff		
c) Country partner - government		
d) Other donor (as in joint evaluation)		
e) Beneficiaries – farmers, small enterprises, households		
f) Others who participated (describe or paste in below)		
a) Unable to determine		
26. Participatory – phase of evaluation (check all that apply)		
b) Evaluation design/methods selection		
c) Data collection		
d) Data analysis		
e) Formulation of recommendations		
f) Other type of participation (describe or paste in below)		
g) Unable to determine		
27. Recommendations		
Number of recommendation provided in the report's recommendations section or summary of recommendations.	Enter number	

3. Evaluation Rating Checklist Instrument³⁸

Evaluation Rating Checklist	YES	NO	CNP ³⁹
Executive Summary			
1. Does the Executive Summary accurately reflect the most critical elements of the report?			
Program/Project Background			
2. Are the basic characteristics of the program, project or activity described (title, dates, funding organization, budget, implementing organization, location/map, target group)?			
3. Is the program or project's "theory of change" described (intended results (in particular the project purpose); development hypotheses; assumptions)			
Evaluation Purpose			
4. Does the evaluation purpose identify the management reason(s) for undertaking the evaluation?			
Evaluation Questions			
5. Are the evaluation questions clearly related to evaluation purpose?			
6. Are the evaluation questions in the report identical to the evaluation questions in the evaluation SOW?			
7. If the questions in the body of the report and those found in the SOW differ, does the report (or annexes) state that there was written approval for changes in the evaluation questions?			
Methodology			
8. Does the report (or methods annex) describe <u>specific</u> data collection methods the team used?			
9. Are the data collection methods presented (in the report or methods annex) in a manner that makes it clear which specific methods are used to address <u>each</u> evaluation question? (e.g., matrix of questions by methods)			
10. Does the report (or methods annex) describe <u>specific</u> data analysis methods the team used? (frequency distributions, cross-tabulations; correlation; reanalysis of secondary data)			
11. Are the data analysis methods presented (in the report or methods annex) in a manner that makes it clear how they are associated with the evaluation questions or specific data collection methods?			
Team Composition			
12. Did the report (or methods annex) indicate that the evaluation team leader was external to USAID?			
13. Did the report (or methods annex) identify at least one evaluation specialist on the team?			
14. Did the report (or methods annex) identify local evaluation team members?			
15. Did the report indicate that team members had signed Conflict of			

³⁸ Two items, numbers 21 and 36, were removed from the checklist after it was determined through inter-rater reliability checks that these elements produced inconsistent and unreliable results. These elements looked at (a) the association of findings to the data sources from which they came, and (b) the inclusion of each and every data collection instrument in annexes. Further explanations of their removal can be found in the body of the report.

³⁹ Conditions required to answer the question are not all present.

Evaluation Rating Checklist	YES	NO	CNP ³⁹
Interest forms or letters? <i>(check if the report says this or the COI forms are included in an annex)</i>			
Study Limitations			
16. Does the report include a description of study limitations (lack of baseline data; selection bias as to sites, interviewees, comparison groups; seasonal unavailability of key informants)?			
Responsiveness to Evaluation Questions			
17. Is the evaluation report structured to present findings in relation to evaluation questions, as opposed to presenting information in relation to program/project objectives or in some other format?			
18. Are all of the evaluation questions, including sub-questions, answered primarily in the body of the report (as opposed to in an annex)?			
19. If any questions were not answered, did the report provide a reason why?			
Findings			
20. Did the findings presented appear to be drawn from social science data collection and analysis methods the team described in its study methodology (including secondary data it assembled or reanalyzed)?			
22. In the presentation of findings, did the team draw on data from the range of methods they used rather than answer using data from primarily one method?			
23. Are findings clearly distinguished from conclusions and recommendations in the report, at least by the use of language that signals transitions (“the evaluation found that....”, “the team concluded that”)?			
24. Are quantitative findings reported precisely, i.e., as specific numbers or percentages rather than general statements like “some”, “many”, or “most”?			
25. Does the report present findings about unplanned/unanticipated results?			
26. Does the report discuss alternative possible causes of results/outcomes it documents?			
27. Are evaluation findings disaggregated by sex at all levels (activity, outputs, outcomes) when data are person-focused?			
28. Does the report explain whether access/ participation and/or outcomes/benefits were different for men and women when data are person-focused?			
Recommendations			
29. Is the report’s presentation of recommendations limited to recommendations? <i>(free from repetition of information already presented or new findings not previously revealed)</i>			
30. Do evaluation recommendations meet USAID policy expectations with respect to being specific? <i>(states clearly what is to be done, and possibly how?)</i>			
31. Do evaluation recommendations meet USAID policy expectations with respect to being directed to a specific party? <i>(identifies who should do it)</i>			
32. Are all the recommendations supported by the findings and conclusions presented? <i>(Can a reader can follow a transparent path from findings to conclusions to recommendations?)</i>			

Evaluation Rating Checklist	YES	NO	CNP ³⁹
Annexes			
33. Is the evaluation SOW included as an annex to the evaluation report?			
34. Are sources of information that the evaluators used listed in annexes?			
35. Are data collection instruments provided as evaluation report annexes?			
37. Were any “Statements of Differences” included as evaluation annexes (prepared by team members, the Mission, the Implementing Partner, or other stakeholder)?			
Evaluation Data Warehousing			
38. Does the evaluation report explain how the evaluation data will be transferred to USAID (survey data, focus group transcripts)?			
SOW Leading Indicator of Evaluation Quality (answer if SOW is a report annex)			
39. Does the evaluation SOW include a copy or the equivalent of Appendix I of USAID’s evaluation policy?			

4. Short Form: Evaluation Ten Point Score Elements

Executive Summary	
1. Is there an Executive Summary which accurately reflects the most critical elements of the report?	
Program/Project Background	
2. Are the basic characteristics and “theory of change” of the program, project or activity described (title, dates, funding organization, budget, implementing organization, location/map, target group)?	
Methodology	
3. Does the report (or methods annex) describe <u>specific</u> data collection methods the team used?	
4. Does the report (or methods annex) describe <u>specific</u> data analysis methods the team used?	
Study Limitations	
5. Does the report include a description of study limitations (lack of baseline data; selection bias as to sites, interviewees, comparison groups; seasonal unavailability of key informants)?	
Findings	
6. Did the findings presented appear to be drawn from social science data collection and analysis methods the team described in its study methodology (including secondary data it assembled or reanalyzed)?	
7. Are findings clearly distinguished from conclusions and recommendations in the report, at least by the use of language that signals transitions (“the evaluation found that.....”, “the team concluded that”)?	
Recommendations	
8. Are all the recommendations supported by the findings and conclusions presented? (Can a reader follow a transparent path from findings to conclusions to recommendations?)	
Annexes	
9. Is the evaluation SOW included as an annex to the evaluation report?	
10. Are data collection instruments provided as evaluation report annexes?	

5. Basic Evaluation Characteristics Descriptions Instrument Handbook

1.	<p><u>What kind of document is it?</u> The purpose of this question is to identify when documents are miscoded in the DEC. It is not uncommon to find documents such as pre-project assessments, GAO or IG audits, or evaluation guides, among other documents, mixed in with actual evaluations. Please indicate which of the available options the document you are coding falls under and provide a description if “other.” If for some reason you are unable to determine what kind of document it is, please let the activity leader know.</p> <p><u>IF NOT AN EVALUATION STOP HERE AND MOVE ON TO THE NEXT EVALUATION ASSIGNED TO YOU!</u></p>
2.	<p><u>Year Published</u> – This information was included on the spreadsheet provided to you and represents how it was entered in the DEC. Please confirm if the information is accurate by comparing it to the year indicated in the report, usually on the cover page or inside cover. If incorrect, provide the correct information.</p>
3.	<p><u>Month Published</u> – This information was not included in the spreadsheet provided, but will be important for splitting up some years, such as 2001 to fully capture when the evaluation policy would have taken effect. Both the month and year should be visible on the front cover or inside cover of the report. Please use the dropdown list provided to select the appropriate month</p>
4.	<p><u>Document Title</u> - This information was included on the spreadsheet provided to you and represents how it was entered in the DEC. Please confirm if the information is accurate by comparing it to the title on the cover page of the report. If the title is abbreviated either in the spreadsheet or in the report, and you are certain you are reading the right report, you do not need to correct the wording. Please confirm by indicating “yes” and move on to the next item. If incorrect, please indicate “no” and provide the correct title.</p>
5.	<p><u>Authoring Organization</u> - This information was included on the spreadsheet provided to you and represents how it was entered in the DEC. Please confirm if the information is accurate by comparing it to the information provided in the report, usually on the cover page or inside cover but perhaps in the body of the report. If the information is accurate, pick “yes” and if the information is incorrect, pick “no” and then enter the correct information.</p>
6.	<p><u>Sponsoring Organization</u> - This information was included on the spreadsheet provided to you and represents how it was entered in the DEC. Please confirm if the information is accurate by comparing it to the information provided in the report, this may be buried in the body of the report. We are looking for the information to be as specific as possible. If “USAID/Georgia” is possible then “USAID” is insufficient. Additionally, there may be more than one sponsoring organization provided. If this is the case, please provide all sponsoring organizations listed separated by a semicolon. If the information is accurate, pick “yes” and if the information is incorrect, pick “no” and then enter the correct information.</p>
7.	<p><u>Geographic Descriptor</u> - This information was included on the spreadsheet provided to you and represents how it was entered in the DEC. Please confirm if the information is accurate by comparing it to the geographic focus of the report as mentioned in the introduction or perhaps title. If the information is accurate, pick “yes” and if the information is incorrect, pick “no” and then enter the correct information.</p>
8.	<p><u>Primary Subject</u> - This information was included on the spreadsheet provided to you and represents</p>

	how it was entered in the DEC. Please confirm if the information is accurate by comparing it to the general subject matter of the project being evaluated. If the information is accurate, pick “yes” and if the information is incorrect, pick “no” and then enter the correct information.
9.	<u>USAID Evaluation Activity Manager</u> – For every evaluation taking place there is a USAID representative assigned to manage the evaluation, often coming from the program office. This person is not always identified, but when they are we need to capture that information. The most likely place will be on the acknowledgements page or on one of the cover pages. It may also be referenced in the methodology section. Please be critical and recognize that being the COR is not sufficient in and of itself; an activity manager may be the COR, the COR is not always the activity manager. If present, please provide their name.
10.	<u>Report Length</u> – This item has two parts <ol style="list-style-type: none"> Executive Summary: Please provide the exact number of pages of the executive summary. If there is only one line on a fifth page it counts as five pages Evaluation Report: This refers to the entire evaluation report including the executive summary, but excluding the annexes or cover pages. Begin your count when the narrative text begins. Please provide the exact number of pages of the evaluation report. If there is only one line on a twenty-fifth page it counts as twenty-five pages.
11.	<u>Evaluation Type</u> - Evaluation type can include an impact evaluation, performance evaluation, or a hybrid of the two. Please refer to the Evaluation Policy (box 1 page 2) for specific definitions of impact and performance evaluations. A hybrid evaluation must include both performance and impact questions and must include a design with two parts, one that establishes at the counterfactual and one that does not. Please choose the appropriate evaluation type from the dropdown menu. If you are unable to determine, pick that option.
12.	<u>Timing</u> – This item is identifying when the evaluation is taking place in relation to the project/program being evaluated. The options include during implementation (at a specific point during the project/program, e.g., in year 2 of 4), approaching the end of a project/program (e.g., in the final year of a long intervention or in the last months of a shorter evaluation), continuous (e.g., for an impact evaluation where the intervention is evaluated throughout its life cycle), or ex-post (evaluation started – not just published – any time from immediately after to several years after project close-out). Please choose the appropriate evaluation timing from the dropdown menu. If you are unable to determine, pick that option.
13.	<u>Scope</u> – This item refers to what exactly was being evaluated. Evaluations can look at individual projects or can look at multiple projects at a time and they can focus on an individual country or a group of countries. It is important for our purposes to be able to distinguish evaluations based on their scope. Some of the scopes provided are fairly straightforward while others are a bit more nuanced and are given more detail below. When evaluating multiple projects within a given country there are three options <ul style="list-style-type: none"> A program-wide evaluation would explicitly examine every element within one of the country mission’s Development Objectives (DOs). DOs focus on large technical issues such as economic growth or food security and would encompass all elements that contribute to achieving the DO. A sector-wide evaluation would look at all, or a sample of, the projects within a given

	<p>technical sector such as agriculture or education.</p> <ul style="list-style-type: none"> The category “other multi-project single-country” might focus on all, or a sample of, the projects within a geographic region of a country or a group of activities, for example, focused on youth employment. <p>When evaluating projects or programs across multiple countries, there are four options</p> <ul style="list-style-type: none"> An example of a single-project multi-country evaluation might focus on an approach to dealing with sexual violence in schools in Malawi and Ghana An example of a multi-project multi-country evaluation might focus on a sample of Mission-funded trade projects around the world A regional program or project evaluation is one that is funded by a regional office or bureau and is focused on a specific geographic region or group of countries. For example, climate change along the Mekong River. A global project is funded through USAID/Washington. For example, a project that can help any mission do a gender assessment. <p>Please choose the appropriate evaluation scope from the dropdown menu. If you are unable to determine, pick that option.</p> <p>If sufficient information is provided, but you are not confident in identifying the scope, please contact the team leader and activity manager for assistance.</p>
14.	<p><u>Evaluation Purpose (policy)</u> – The evaluation policy is clear about having two reasons to conduct evaluations at the agency level. The first is accountability (measuring project effectiveness, relevance and efficiency, disclosing those findings to stakeholders, and using evaluation findings to inform resource allocation and other decisions) and the second is learning (Evaluations of projects that are well designed and executed can systematically generate knowledge about the magnitude and determinants of project performance, permitting those who design and implement projects, and who develop programs and strategies – including USAID staff, host governments and a wide range of partners – to refine designs and introduce improvements into future efforts). An evaluation purpose can only count in this category if explicitly stated as such, using the words “learning” or “accountability,” though it is also acceptable to use both. If neither of these words is explicitly stated pick that option.</p>
15.	<p><u>Evaluation Purpose (management)</u> – The management purpose of the evaluation must be explicit in regards to the decisions and actions the evaluation is intended to inform and should come from the body of the evaluation if possible before taking from the executive summary, but should not be taken from the SOW. An evaluation can have more than one management purpose. Response options based on the most common management purposes from previous studies are shown on the demographic sheet. Please indicate all options that apply by choosing “yes” or “no” for each option using the dropdown list provided. If you found a management purpose other than one of the options provided, please pick yes for the “other” option and paste the language into the space provided. If you were not able to identify a management purpose from any of the options provided, pick yes on the final option “unable to determine.”</p> <p>Be sure you put either yes or no for every option in this set</p>
16.	<p><u>What was the evaluation asked to address</u> – Answer options for this question include: questions, issues, and other. For this item, identify what the evaluation team stated that they were asked to address in the evaluation. Please look in the body of the report for this item, and if no information is</p>

	<p>available there then look in the evaluation SOW. The two most likely responses will be questions or issues. USAID policy and supporting documents are requiring the use of questions, but it is not uncommon to find issues instead. If an evaluation team claims to be asked to address something other than questions or issues, please check “other” and include the language used in the report. If there is no language in the report, or in the SOW, on what the evaluation team was asked to address, please choose that option. If issues or anything other than questions are indicated please skip forward to Q18.</p>
17.	<p><u>Number of Evaluation Questions</u> – Complete this section only if you answered “questions” on 16, above. This section includes four elements.</p> <ol style="list-style-type: none"> Are the questions numbered? This is a yes/no question about whether questions (not issues) found in the body of the report, or in the SOW if there were none in the body of the report, had been assigned numbers. If there are questions in both the body of the report and the SOW, the questions in the body of the report take precedence in terms of answering all elements of this set of questions. To how many questions were full numbers assigned and what is the total of those numbers? In the simplest instance, questions would be numbered 1-5. If there are sub-questions, (e.g., 5a, 5b) then the highest number of questions would still be 5. In other instances, questions might be in groups (e.g., A, 1-5, and then B, 1-6). In this type of case the number of numbered questions would be 11. If you answered “no” on 17 (a) above, enter 0 (zero) for 17 (b) How many questions marks were included among the questions? This is a simple count of how many question marks were used in presenting the questions in the body of the report, or in the SOW if no questions were found in the body of the report. Don’t worry about hidden or compound questions, just count question marks. If there are questions with no question marks, they cannot be counted, only questions with question marks. How many total questions, including compound (hidden) questions? For this item, we are looking for a count of all questions beyond those distinguished by a question mark. Compound, or hidden questions, are questions with an “and” in them or perhaps a list of items an evaluator is being asked to look at within a specific question. An example of this might be, “what was the yield and impact for each crop variety?”
18.	<p><u>Evaluation Design/Approach to Causality/Attribution Included</u> – If the evaluation team is responsible for answering one or more questions or issues that ask about causality or attribution pick “yes” and move to the next item (#18). If there is no question or issue asking about causality or attribution, pick “no” and move on to item 19.</p>
19.	<p><u>Evaluation Design Types</u> – For questions or issues of causality and attribution, there are three categories of evaluation designs to choose from. In order to fall into one of these categories the evaluation design must be specifically discussed in the body of the evaluation report and not exclusively in an annex. If not discussed, or if discussed exclusively in an annex exclusively, please pick yes for the final option “design not presented.” If a design was discussed, please indicate which of the following three design categories it falls into and provide the page number where it can be found in the report.</p> <ul style="list-style-type: none"> Experimental design – this type of design will only be used for impact evaluations and might be referenced using one of the following keywords: experimental design, control group, randomized assignment, or randomized controlled trial.

	<ul style="list-style-type: none"> Quasi-experimental design – this type of design will only be used for impact evaluations and might be referenced using one of the following keywords: quasi-experimental, comparison group, propensity score matching, interrupted time series, or regression discontinuity. Non-experimental design – a design in this category uses an approach examining causality/attribution that does not include an experiment. Terminology associated with one of these designs might include language identifying and eliminating alternative possible causes (modus operandi), outcome mapping, action research, contribution analysis, or case study.
20.	<p><u>Data Collection Methods (team said it planned to use)</u> – For this item, we are looking for every data collection method that the evaluation team stated that they planned to use (either in the body of the report or in a methodology annex). In the instance that the data collection team introduces a data collection method, but mis-states what the method actually is, and there is enough information provided for you as a coder to appropriately re-categorize it, please do so (e.g., if an evaluation claims to be doing quantitative interviews, but the description and a look at the data collection instrument indicate that it is actually a survey, mark it as a survey). An evaluation can use more than one data collection method. A list of data collection methods based on the most common methods used in previous studies are shown on the demographic sheet. Note that document reviews/secondary data sources reflect all non-project related documents while project documents refer to any project related documents including, but not limited to PMPs, Quarterly Reports, etc. Please indicate all options that apply by choosing “yes” or “no” for each option using the dropdown list provided. If you found a data collection method other than one of the options provided, please pick yes for the “other” option and paste the language into the space provided. If a data collection method is insufficiently detailed enough to fit into an option provided (i.e., “interviews” and not “key-informant interviews” or “other interviews”) then check “other” and in the area provided indicate “interviews – not specified.” If you were not able to identify a data collection method from any of the options provided, pick yes on the final option “unable to determine.”</p> <p>Be sure you put either yes or no for every option in this set</p>
21.	<p><u>Data Collection Methods (data actually used)</u> - For this item, we are looking for the presentation of data that shows which data collection methods were actually used. For example, “20% of the survey respondents said” indicates that the survey method was actually used. The demographic sheet shows the same list of data collection methods as you saw in item 19. Again, note that document reviews/secondary data sources reflect all non-project related documents while project documents refer to any project related documents including, but not limited to PMPs, Quarterly Reports, etc. For every method you mark that they planned to use, look to see if there was data linked to words about the method that would indicate it was actually used. Additionally, for any data linked to methods that were used but which you did not code as methods they stated they planned to use, mark “yes” for that data collection method. In the instance that the data collection team introduces a data collection method, but mis-states what the method actually is, and there is enough information provided for you as a coder to appropriately re-categorize it, please do so (e.g., if an evaluation claims to be doing quantitative interviews, but the description and a look at the data collection instrument indicate that it is actually a survey, mark it as a survey).</p> <p>Please indicate all options that apply by choosing “yes” or “no” for each option using the dropdown list provided. If you found a data collection method other than one of the options provided, please pick yes for the “other” option and paste the language into the space provided. If you were not able to identify a data collection method from any of the options provided, pick yes on the final option “unable to determine.”</p>

	Be sure you put either yes or no for every option in this set																																				
22	<p>Data Analysis Methods (team said it planned to use) – For this item, we are looking for every data analysis method that the evaluation team stated that they planned to use (either in the body of the report or in a methodology annex). An evaluation can use more than one data analysis method. A list of data analysis methods based on the most common methods used in previous studies are shown on the demographic sheet. An additional option for noting where the team described how it planned to synthesize data from multiple methods (mixed methods) is also shown on the demographic sheet. Please indicate all options that apply by choosing “yes” or “no” for each option using the dropdown list provided. If you found a data analysis method other than one of the options provided, please pick yes for the “other” option and paste the language into the space provided. If you were not able to identify a data analysis method from any of the options provided, pick yes on the final option “unable to determine.”</p> <p>Be sure you put either yes or no for every option in this set</p>																																				
23	<p>Data Analysis Methods (data actually used) - For this item, we are looking for the presentation of data that shows which data analysis methods were actually used. Examples of the kinds of language you might find if they used particular methods can be found in the table below. The demographic sheet shows the same list of data analysis methods as you saw in item 21. For every method you mark that they planned to use, look to see if there was analysis language, tables, or graphs that would indicate it was actually used. Additionally, for any analyses that were used but which you did not code as analyses they stated they planned to use, mark “yes” for that data analysis method.</p> <p>Please indicate all options that apply by choosing “yes” or “no” for each option using the dropdown list provided. If you found a data analysis method other than one of the options provided, please pick yes for the “other” option and paste the language into the space provided. If you were not able to identify a data analysis method from any of the options provided, pick yes on the final option “unable to determine.”</p> <p>Be sure you put either yes or no for every option in this set</p>																																				
	<table border="1"> <thead> <tr> <th>Q.21 They Said They Plan to Do</th><th>Q.22 They Show They Did</th></tr> </thead> <tbody> <tr> <td>Descriptive Statistics</td><td></td></tr> <tr> <td>Frequency</td><td>Question 28: 23 said yes; 7 said no</td></tr> <tr> <td>Percentage</td><td>77% of respondents said “yes”</td></tr> <tr> <td>Ratio</td><td>The ratio of books to students is 1:6</td></tr> <tr> <td>Cross-tabulation</td><td> <table border="1"> <thead> <tr> <th>Loan Status</th><th>Men</th><th>Women</th><th>Total</th></tr> </thead> <tbody> <tr> <td>Took a loan</td><td>16</td><td>8</td><td>24</td></tr> <tr> <td>Didn’t take a loan</td><td>8</td><td>16</td><td>24</td></tr> <tr> <td>Total</td><td>24</td><td>24</td><td>48</td></tr> </tbody> </table> </td></tr> <tr> <td>Inferential Statistics</td><td></td></tr> <tr> <td>Correlation (tells how closely related two variables are)</td><td>Correlation coefficient; statistical significance</td></tr> <tr> <td>Regression</td><td>Regression coefficient; statistical significance</td></tr> <tr> <td>t-test (compares averages for groups)</td><td>Difference between means; t value; statistical</td></tr> </tbody> </table>	Q.21 They Said They Plan to Do	Q.22 They Show They Did	Descriptive Statistics		Frequency	Question 28: 23 said yes; 7 said no	Percentage	77% of respondents said “yes”	Ratio	The ratio of books to students is 1:6	Cross-tabulation	<table border="1"> <thead> <tr> <th>Loan Status</th><th>Men</th><th>Women</th><th>Total</th></tr> </thead> <tbody> <tr> <td>Took a loan</td><td>16</td><td>8</td><td>24</td></tr> <tr> <td>Didn’t take a loan</td><td>8</td><td>16</td><td>24</td></tr> <tr> <td>Total</td><td>24</td><td>24</td><td>48</td></tr> </tbody> </table>	Loan Status	Men	Women	Total	Took a loan	16	8	24	Didn’t take a loan	8	16	24	Total	24	24	48	Inferential Statistics		Correlation (tells how closely related two variables are)	Correlation coefficient; statistical significance	Regression	Regression coefficient; statistical significance	t-test (compares averages for groups)	Difference between means; t value; statistical
Q.21 They Said They Plan to Do	Q.22 They Show They Did																																				
Descriptive Statistics																																					
Frequency	Question 28: 23 said yes; 7 said no																																				
Percentage	77% of respondents said “yes”																																				
Ratio	The ratio of books to students is 1:6																																				
Cross-tabulation	<table border="1"> <thead> <tr> <th>Loan Status</th><th>Men</th><th>Women</th><th>Total</th></tr> </thead> <tbody> <tr> <td>Took a loan</td><td>16</td><td>8</td><td>24</td></tr> <tr> <td>Didn’t take a loan</td><td>8</td><td>16</td><td>24</td></tr> <tr> <td>Total</td><td>24</td><td>24</td><td>48</td></tr> </tbody> </table>	Loan Status	Men	Women	Total	Took a loan	16	8	24	Didn’t take a loan	8	16	24	Total	24	24	48																				
Loan Status	Men	Women	Total																																		
Took a loan	16	8	24																																		
Didn’t take a loan	8	16	24																																		
Total	24	24	48																																		
Inferential Statistics																																					
Correlation (tells how closely related two variables are)	Correlation coefficient; statistical significance																																				
Regression	Regression coefficient; statistical significance																																				
t-test (compares averages for groups)	Difference between means; t value; statistical																																				

	<p>with continuous variables, like money)</p> <p>Chi-square (compares answers for groups with discontinuous variables (high, medium, low)</p> <p>Content Analysis</p> <p>Code key words, phrases, concepts mentioned in open-ended questions, group interviews or focus groups; identify dominant patterns, or quantify the results of pattern coding</p>	<p>significance</p> <p>Difference between groups; statistical significance</p> <p>Discussion of dominant content or patterns of responses to open-ended (qualitative, or transformed into quantitative form)</p>	
24.	<p><u>Participatory Mentioned?</u> For this item, if there was any mention of a participatory method or approach then it counts even if there is no further discussion of who participated or in which phase they participated.</p> <p>If yes, indicate who participated (beyond contributing data) and at what stage of the evaluation in questions 24 and 25 below. If not, please skip questions 24 and 25.</p>		
25.	<p><u>Participatory (when)</u> – There are various stages at which people outside of the evaluation team may become involved in the evaluation. We are looking to identify participation at any of the stages that an evaluation report indicates that it occurred. Note that if a person is on the evaluation team, even if a country national, USAID staff, or implementing partner staff, they cannot be considered as participating in the evaluation for this item.</p> <p>Please indicate all options that apply by choosing “yes” or “no” for each option using the dropdown list provided. If you found a stage or type of participation other than one of the options provided, please pick yes for the “other” option and paste the language into the space provided. If you were able to determine that participation took place but not at what particular stage of the process, pick yes on the final option “unable to determine.”</p>		
26.	<p><u>Participatory (who)</u> – There are various groups of people outside of the evaluation team who may become involved in the evaluation. Such groups could include, but are not limited to, USAID representatives (other than the evaluation activity manager), project/program implementing partners including the government, other donors, or beneficiaries. Note that if a person is on the evaluation team, even if a country national, USAID staff, or implementing partner staff, they cannot be considered as participating in the evaluation for this item. Please indicate all options that apply by choosing “yes” or “no” for each option using the dropdown list provided. If you identified stakeholders who participated in the evaluation process other than one of the options provided, please pick yes for the “other” option, and paste the language into the space provided. If you were able to determine that participation took place but not who participated, pick yes on the final option “unable to determine.”</p>		
27.	<p><u>Recommendations</u> – Please provide the number of recommendations provided in a recommendations section, or a summary of recommendations in the body of the report, and not in an executive summary. Count the number of identifiable recommendations, whether they are shown as numbers, letters, or bullets. [This is a change in instructions. Do not look inside the bullets or numbered recommendations to separate out where they are compound in nature. Simply count what the evaluation calls recommendations.]</p>		

6. Evaluation Rating Checklist Instrument Handbook⁴⁰

Please be aware that CNP means Conditions Not Present and indicates insufficient information exists to answer a checklist item. Be aware of when a CNP response is possible. Also, check for relationships between questions (i.e., if Q9 is “no” then Q10 must be “CNP”).

Executive Summary	
1. Does the Executive Summary present an accurate reflection of the most critical elements of the report?	An Executive Summary must provide an accurate representation of the gist of the evaluation report without adding any new “material” information or contradicting the evaluation report in any way. “Critical” implies that not all information included in the evaluation report needs to be present in the executive summary, but that critical information from <u>all major elements should be discussed (i.e., evaluation purpose, questions, background information, methods, study limitations, findings, and recommendations)</u> . If an executive summary is not present, check “CNP” in the dropdown box provided. If the executive summary is consistent with the language that you see here check “yes.” If it does not conform to what you see here, check “no.”
Program/Project Background	
2. Are the basic characteristics of the project or program described (title, dates, funding organization, budget, implementing organization, location/map, target group)?	The project description plays a critical role in enabling the reader to understand the context of the evaluation, and involves several characteristics such as the title, dates, funding organization, budget, implementing organization, location/map, and target group. While every one of these characteristics plays an important role and should be present, for the purposes of this study we are looking for a holistic view of whether the project is sufficiently described. If one or two characteristics are missing or weak but you get the gist of the project and can answer all future questions, then check “yes.” If not, check “no.”
3. Is the project or program’s “theory of change” described (intended results (in particular the project Purpose); development hypotheses; assumptions)	The “theory of change” describes, via narrative or graphic depiction, the intended results and causal logic that explains how they will be achieved. You may see this described as the development hypotheses and assumptions underlying the project or program. We are looking for the theory of change to be presented fully in one place before the introduction of findings. If these elements of a theory of change are present, even if weak, check “yes.” If they are not present, check “no.”
Evaluation Purpose	
4. Does the evaluation purpose identify the	Evaluation policy states that USAID is conducting evaluations for learning and accountability purposes. Beyond that, it is important

⁴⁰ Two items, numbers 21 and 36, were removed from the checklist after it was determined through inter-rater reliability checks that these elements produced inconsistent and unreliable results. These elements looked at (a) the association of findings to the data sources from which they came, and (b) the inclusion of each and every data collection instrument in annexes. Further explanations of their removal can be found in the body of the report.

management reason(s) for undertaking the evaluation?	that the evaluation purpose identifies the specific decisions or actions the evaluation is expected to inform (e.g., continue, terminate, expand, redesign). If an evaluation purpose is not present, or is only present in the SOW, check “CNP” in the dropdown box provided. If the evaluation purpose describes specific decisions or actions the evaluation will inform, consistent with those illustrated here, check “yes.” If it does not conform to what you see here, check “no.”
Evaluation Questions	
5. Are the evaluation questions clearly related to the evaluation purpose?	The evaluation questions, or issues, as stated in the evaluation report should have a direct and clear relationship with the purpose of the evaluation in order to be effective and useful (i.e., learning, accountability, upcoming management decisions). If no evaluation questions/issues are provided in the body of the report before the findings, or in the SOW, check “CNP” in the dropdown list provided (questions from the SOW are acceptable). If the evaluation questions/issues are related to the evaluation purposes stated in the report, check “yes.” If not, check “no.”
6. Are the evaluation questions in the report identical to the evaluation questions in the evaluation SOW?	This question is about evaluation questions found in the body of the report and in the SOW. There must be questions in both places in order address this question. If questions are present in only one of these two places, mark CNP.
7. If the questions in the body of the report and those found in the SOW differ, does the report (or annexes) state that there was written approval for changes in the evaluation questions?	As the evaluation SOW is essentially the contract from which evaluators are working from, it is imperative that the questions/issues they list, and address, in the evaluation report before the presentation of findings, match those included in the SOW word for word. If, for some reason the evaluation team changed, removed, or added evaluation questions/issues, they could only have done so with written approval. While this written approval does not need to be included in an annex, it does need to be mentioned in the body of the report. If the answer to 6a is “yes” or “CNP” then mark 6b as “CNP.” If the answer to 6a is “no” then answer 6b with a “yes” or “no.”
Methodology	
8. Does the report (or methods annex) describe <u>specific</u> data collection methods the team used?	USAID requires that an evaluation report identify the data collection methods that were used, but does not indicate where this information must be presented. It is not uncommon for a 1-3 page methodology description to be included in the body of the report with a longer and more detailed methods section provided as an annex. To count for our purposes, the methods description must be specific on how data will be collected. It is insufficient to simply say, “interviews will be conducted,” and instead must provide detailed information on the kinds of interviews (key informant, individual, group, or focus group), the number of interviews, and who was interviewed (specific names). If a description of data

	collection methods is provided at a level of detail equivalent to that illustrated here was found, check “yes.” If not, check “no.”
9. Are the data collection methods presented (in the report or methods annex) in a manner that makes it clear which specific methods are used to address <u>each</u> evaluation question (e.g., matrix of questions by methods)?	It is useful for the evaluation report to make it clear which of the data collection methods described were used to gather data to answer each specific evaluation question/issue, including all sub-questions/issues. This information may be available within the body of the report or may be found in a methods or design annex. While the methods can be associated to questions in a variety of ways, a good example would be a “method x question” matrix. If no data collection methods are provided, or if no questions/issues exist, check the box for “CNP.” If specific methods are identified on a question-by-question (or issue-by-issue) basis, check “yes.” If not, check “no.”
10. Does the report (or methods annex) describe <u>specific</u> data analysis methods the team used? (frequency distributions; cross-tabulations; correlation; reanalysis of secondary data)	USAID requires that an evaluation report identify the data analysis methods that were used, but does not indicate where this information must be presented. It is not uncommon for a 1-3 page methodology description to be included in the body of the report with a longer and more detailed methods section provided as an annex. To count for our purposes, the data analysis methods description must be <u>specific about how, or by using what method, data will be analyzed</u> . It is insufficient to simply say, “qualitative and quantitative analyses will be conducted” and instead must provide detailed information on the kinds of analyses to be conducted (e.g., frequency distributions, cross-tabs, correlations, on the quantitative side, or content analysis, pattern analysis or some other type of qualitative analysis they used). If a description of data analysis methods is provided at a level of detail equivalent to that illustrated here was found, check “yes.” If not, check “no.”
11. Are the data analysis methods presented (in the report or methods annex) in a manner that makes it clear how they are associated with the evaluation questions or specific data collection methods?	It is useful for the evaluation report to make it clear which of the data analysis methods described were used to analyze data to answer evaluation questions/issues or to analyze data from specific methods. [The question parallels Question 8 above for data collection methods.] Information on data analysis methods may be available within the body of the report or may be found in a methods or design annex. While the methods can be associated to questions/issues in a variety of ways, a good example would be a “method x question” matrix. If no data analysis methods are provided (marked “no” for previous question, #9), or if no questions exist, check the box for “CNP.” If specific methods are identified on a question-by-question (or issue-by-issue) basis, check “yes.” If not, check “no.”
Team Composition	
12. Did the report (or methods annex) indicate that the evaluation team leader was external to	USAID counts an evaluation as being external if the team leader is external, meaning that the team leader is an independent expert from outside of USAID who has no fiduciary relationship with the implementing partner. If the evaluation is a self-evaluation

USAID?	(USAID is evaluating their own project) then this answer must be no. For our purposes, the evaluation must indicate the team leader in either the body of the report (including cover or title page) or in the methods section. Please look to the body of the report first and then move on to the methods annex. A simple “search” function for the term “leader” may expedite this process. <u>If the report is not explicit in stating the team leader was external, it may be inferred from a description of the team leader or the organization with which they are associated (e.g., university professor or evaluation firm that is not the project implementer).</u> Independence may also be confirmed via a “no-conflict of interest” statement often included as an annex. If the report identifies that the team was independent, but there is no designated team leader, check “CNP.” If the report identifies an external evaluation team leader, check “yes.” If not, check “no.”
13. Did the report (or methods annex) identify at least one evaluation specialist on the team?	At least one member of the evaluation team must be an evaluation specialist and clearly indicated as such in either the body of the report or in the methods annex. The term “evaluation specialist” must be explicit and not implied. If at least one evaluation specialist is identified then check “yes.” If not, check “no.”
14. Did the report (or methods annex) identify local evaluation team members?	USAID encourages the participation of country nationals on evaluation teams. The report need not use the word “local” specifically, but can be referred to by designation such as “Brazilian education specialist,” if in Brazil. This person could be any country national, including a foreign service national (FSN). Simply guessing a person’s country of origin based on their name is insufficient. If the report identified one or more country nationals on the evaluation team, check “yes.” If not, check “no.”
15. Did the report indicate that team members had signed Conflict of Interest forms or letters (check if the report says this or the COI forms are included in an annex)?	USAID requires that evaluation team members certify their independence by signing statements indicating that they have no conflict of interest or fiduciary involvement with the project or program they will evaluate. USAID guidance includes a sample Conflict of Interest form. It is expected that an evaluation will indicate that such forms, or their equivalent, are on file and available or are provided in an evaluation annex. If the evaluation states that they are on file or provides them in an annex, check “yes.” If not, check “no.”
Study Limitations	
16. Does the report include a description of study limitations (lack of baseline data; selection bias as to sites, interviewees, comparison groups; seasonal unavailability	It is quite common for evaluators to run into unexpected interferences with anticipated study designs such as unavailability of key informants or lack of access to activity sites or information needed to implement a sampling plan. In other instances, stakeholder preferences may introduce selection biases or beneficiaries may have already begun reacting to an activity prior to its start. In any such instance, the evaluators are obligated to include these “study limitations” and a description of the impact

of key informants)?	that they would have had on the evaluation. Study limitations may only be included for this item if they directly impact the evaluator's ability to credibly and effectively answer an evaluation question (i.e., if all data can still be collected, even if inconveniently or at a higher cost, it is not a limitation). Limitations do not need to have their own distinct section provided they are located towards the end of the methodology description and before the introduction of findings. If a description of study limitations is included, then check "yes." If not, check "no."
Responsiveness to Evaluation Questions	
17. Is the evaluation report structured to present findings in relation to evaluation questions, as opposed to presenting information in relation to project objectives or in some other format?	The most straightforward way to meet USAID's requirement that every evaluation question/issue be addressed, is a question-by-question (or issue-by-issue) report structure. Historically, evaluations have not always taken this approach, and instead structured the report around such things as project objectives, or locations. If no evaluation questions/issues around which a report could be structured, check "CNP." If the evaluation questions/issues and the team's answers to those questions/issues are the dominant structure of the report, check "yes." If it was something else, check "no."
18. Are <u>all</u> of the evaluation questions, including sub-questions, answered primarily in the body of the report (as opposed to in an annex)	The purpose of an evaluation report is to provide the evaluators' findings and recommendations on <u>each</u> and <u>every</u> evaluation question. Accordingly, USAID expects that the answers to <u>all</u> evaluation questions/issues, including any sub-questions/issues, will be provided primarily in the body of the report. Answering main questions/issues in the body and sub-questions/issues in an annex is not consistent with USAID expectations. If no evaluation questions/issues are provided (either in the body of the report or in an annex) to which a team could respond, check "CNP." If every question/issue, including sub-questions/issues, was addressed in their report, check "yes." If not, check "no."
19. If any questions were not answered, did the report provide a reason why?	If the answer to question 17a is "yes," mark the answer to 17b as "CNP." If the answer to question 17a is "no," does the evaluation report provide an explanation as to why specific questions were not answered or were answered somewhere other than in the body of the report? If there is an explanation, mark "yes," if there is no explanation, mark "no."
Findings	
20. Did the findings presented appear to be drawn from social science data collection and analysis methods the team described in study methodology (including secondary data	USAID's commitment to evidence-based decision-making is necessitating a shift to stronger and more replicable approaches to gathering data and presenting action recommendations to the agency. The more consistent use of credible social science data collection and analysis methods in evaluations is an important step in that direction (e.g., structured and well documented interviews, observation protocols, survey research methods). If the report did not describe the data collection and analysis methods used, check

assembled or reanalyzed)?	“CNP.”
22. In the presentation of findings, did the team draw on data from the range of methods they used rather than answer using data from or primarily one method?	In addressing this question, please reference only those data collection methods that are identified as having been used in the demographics section of this score-sheet. Of the methods they actually used, we are looking for demonstration of a balanced use of data from all data collection methods indicated in the methodology. If no methodologies were introduced from which they could later be drawn on, check “CNP.”
23. Are findings clearly distinguished from conclusions and recommendations in the report, at least by the use of language that signals transitions (“the evaluation found that...” or “the team concluded that...”)?	As defined by the evaluation policy, evaluation findings are based on facts, evidence, and data...[and] should be specific, concise, and supported by quantitative and qualitative information that is reliable, valid, and generalizable. The presence of opinions, conclusions, and/or recommendations mixed in with the descriptions of findings reduces a finding’s ability to meet USAID’s definition.
24. Are quantitative findings reported precisely, i.e., as specific numbers or percentages rather than general statements like “some,” “many,” or “most”?	When presenting quantitative findings it is important to be precise so that the reader knows exactly how to interpret the findings and is able to determine the accuracy of the conclusions drawn by the evaluators. Precision implies the use of specific numbers and/or percentages as opposed to general statements like “some,” “many,” or “most.” If no potentially quantitative findings are provided, check “CNP.”
25. Does the report present findings about unplanned/unanticipated results?	Though evaluators may be asked to look for unplanned or unanticipated results in an evaluation question, it is not uncommon for evaluators to come across such results unexpectedly. If any such results are found, by request or unexpectedly, it is useful to include such information in the evaluation report. If the report presented findings about unplanned or unanticipated results, check “yes.” If no unplanned or unanticipated results are presented, check “no.”
26. Does the report discuss alternative possible causes of results/outcomes it documents?	Though evaluators may be asked to look for alternative causes of documented results or outcomes in an evaluation question, it is possible for evaluators to come across such potential alternative causes unexpectedly. If any such causes are found, it is important that the evaluators bring such information to the attention of USAID. If the report discusses alternative possible causes of documented outcomes or results, check “yes.” If no alternative causes are discussed, check “no.”
27. Are evaluation findings	The evaluation policy and USAID in general are making a big

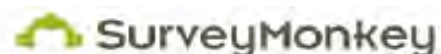
disaggregated by sex at all levels (activity, outputs, outcomes) when data are person-focused?	push for gathering sex-disaggregated data whenever possible. To support this focus of USAID, it is valuable for evaluators to include data collection and analysis methods that enable sex-disaggregation whenever the data they anticipate working with will be person-focused. Such data should be represented at all project levels from activities to outputs to outcomes to the extent possible. If no person-focused data was collected and therefore there was no data that could be disaggregated by sex, check “CNP.”
28. Does the report explain whether access/participation and/or outcomes/benefits were different for men and women when data are person-focused?	USAID expects that evaluations will identify/discuss/explain how men and women have participated in, and/or benefited from, the programs and projects it evaluates. This involves more than simply collecting data on a sex-disaggregated basis. Addressing this issue can be presented in one general section or on a question-by-question basis; either is acceptable. If data was not collected in a person-focused manner for the evaluation, check CNP.
Recommendations	
29. Is the report’s presentation of recommendations limited to recommendations (free from repetition of information already presented or new findings not previously revealed)?	The manner in which recommendations are presented within an evaluation report plays an influential role in the usability of the evaluation report. Recommendations are built off of the information previously introduced through the presentation of findings and conclusions. For this reason, the presentation of recommendations does not need to have the supporting findings and conclusions repeated or any new supporting findings or conclusions introduced. The presence of anything other than specific, practical, and action-oriented recommendations could have a diminishing effect on the usability of the report. If no recommendations are presented in the evaluation report, check “CNP.”
30. Do evaluation recommendations meet USAID policy expectations with respect to being specific (states what exactly is to be done, and possibly how)?	Recommendations that are specific are inherently more actionable than those which are not. The recommendation, “improve management of the project,” is much less specific than one that says “streamline the process for identifying and responding to clinic needs for supplies in order to reduce gaps in service delivery.” If no recommendations are presented in the evaluation report, check “CNP.”
31. Do evaluation recommendations meet USAID policy expectations with respect to being directed to a specific party (identifies who should do it)?	USAID encourages evaluation teams to identify the parties who the evaluation team is saying need take action on each recommendation included in an evaluation report. This feature makes it easier for USAID staff to understand and act on and evaluations implications. If no recommendations are presented in the evaluation report, check “CNP.”

32. Are all the recommendations supported by the findings and conclusions presented (Can a reader can follow a transparent path from findings to conclusions to recommendations)?	Managers are more likely to adopt evaluation recommendation when those evaluations are based on credible empirical evidence and an analysis that transparently demonstrates why a specific recommendation is the soundest course of action. To this end, USAID encourages evaluators to present a clear progression from Findings → Conclusions → Recommendations in their reports, such that none of a report's recommendations appear to lack grounding, or appear out of "thin air." If no recommendations are presented in the evaluation report,
Annexes	
33. Is the evaluation SOW included as an annex to the evaluation report?	This question checks on evaluation team responsiveness to USAID's Evaluation Policy, Appendix 1, requirement for including an evaluation SOW as an evaluation report annex. If the evaluation SOW was not provided as part of the evaluation report, e.g., in an annex, check no. If it was provided, check "yes."
34. Are sources of information that the evaluators used listed in annexes?	This question checks on evaluation team responsiveness to USAID's Evaluation Policy, Appendix 1, requirement for including information about sources of information as an evaluation report annex. Sources include both documents reviewed and individuals who have been interviewed. Generally speaking it is not expected that survey respondents or focus group member names will be individually provided, as these individuals are generally exempted based on common/shared expectations about maintaining confidentiality with respect to individual respondents.
35. Are data collection instruments provided as evaluation report annexes?	This question focuses on the inclusion of data collection instruments in an evaluation annex. If an annex that includes data collection instruments is included with a report or if the report otherwise provides these tools, check "yes." If not, check "no."
37. Were any "Statements of Differences" included as evaluation annexes (prepared by team members, or the Mission, or Implementing Partner, or other stakeholders)	Including "Statements of Differences" has long been a USAID evaluation report option. This question determines how frequently "Statements of Differences" are actually included in USAID evaluations. Statements are often written by evaluation team members, or alternatively by the Mission, a stakeholder, or implementing partner. If one or more "Statements of Differences" are included, check "yes"? If not, check "no."
Evaluation Data Warehousing	
38. Does the evaluation report explain how the evaluation data will be transferred to USAID (survey data, focus group transcripts)?	USAID evaluation policy (p. 10) calls for the transfer of data sets from evaluations to USAID, so that, when appropriate, they can be reused in other assessment and evaluations. Given this requirement, it is helpful if an evaluation report indicates how and when that transfer was made. If the report describes the transfer of evaluation data to USAID, check "yes." If not, check "no"
SOW Leading Indicator of Evaluation Quality (answer if SOW is a report annex)	

<p>39. Does the evaluation SOW include a copy or the equivalent of Appendix 1 of the evaluation policy?</p>	<p>USAID policy requires that statements of work (SOWs) for evaluations include the language of Appendix 1 of the USAID Evaluation Policy. If no SOW is included as an annex to the evaluation report, check “CNP.” If the SOW is included as an annex to the evaluation and includes this language, check “yes.” If not, check “no.”</p>
---	---

Annex D. Evaluation Team Leader Perception Survey Responses



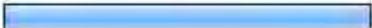
Team Leader Survey







1. Name (optional)

	Response Count
	13
answered question	13
skipped question	12

2. Roughly how many evaluations have you conducted over your career?

		Response Percent	Response Count
One to five evaluations		12.0%	3
Five to ten evaluations		32.0%	8
Ten or more evaluations		56.0%	14
	answered question		25
	skipped question		0

3. How many of those evaluations were USAID evaluations?

		Response Percent	Response Count
All of them		28.0%	7
One to five evaluations		36.0%	9
Five to ten evaluations		8.0%	2
Ten or more evaluations		28.0%	7
	answered question		25
	skipped question		0

4. How many times have you served as a USAID evaluation Team Leader?

Response
Count

23

answered question

23

skipped question

2

5. What is the most recent year in which you served as the Team Leader for a USAID evaluation?

Response
Percent Response
Count

2009

0.0%

0

2010



4.0%

1

2011



20.0%

5

2012



76.0%

19

answered question

25




skipped question

0



6. Please indicate the key factors that you feel have been most instrumental in your being selected to serve as a Team Leader for USAID evaluations?

	Response Percent	Response Count
Evaluation experience (number/types of previous evaluations)	72.0%	18
Sector/topic expertise (health, agriculture, youth, gender)	80.0%	20
Regional or country expertise (South Asia, Malawi)	60.0%	15
Previous service as a USAID staff member, including as a Personal Services Contractor	24.0%	6
Evaluation expertise (evaluation design/methods skills/degrees/experience teaching evaluation)	60.0%	15
Team management expertise/experience	76.0%	19
Other (please specify)		6
answered question		25
skipped question		0



7. When most recently serving as the Team Leader for a USAID evaluation, what choice below best describes your employment status

		Response Percent	Response Count
Staff member of a firm that does evaluations for USAID		8.3%	2
Academic position in a university		8.3%	2
Freelance consultant not affiliated with any institution on a full time basis		83.3%	20
Staff of the implementing partner organization that implemented the project/program		0.0%	0
USAID staff member		0.0%	0
	Other (please specify)		0
answered question			24
skipped question			1




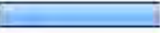


8. As a recent USAID evaluation team leader, were you provided with information about USAID's evaluation quality standards at the start of the evaluation?

		Response Percent	Response Count
Yes		88.0%	22
No		12.0%	3
answered question			25
skipped question			0





9. Who provided you with the information you received on USAID evaluation quality standards?

		Response Percent	Response Count
Staff of the firm or NGO that organized the evaluation team		50.0%	10
USAID staff		50.0%	10
	Other (please specify)		3
answered question			20
skipped question			5

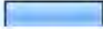



10. Which specific document or documents contained guidance on USAID's evaluation standards? (Check all documents you were given)

		Response Percent	Response Count
An evaluation SOW that included a list of USAID evaluation quality standards		63.6%	14
USAID's Evaluation Policy paper (2011)		63.6%	14
USAID ADS 203		18.2%	4
A USAID TIPS or "How To" Note about Preparing a USAID evaluation		27.3%	6
A USAID evaluation report outline or template		45.5%	10
A USAID checklist for reviewing evaluation reports		45.5%	10
	Other (please specify)		1
answered question			22
skipped question			3




11. How would you characterize the amount of time allocated to the evaluation team for the most recent USAID evaluation for which you were the Team Leader?

		Response Percent	Response Count
More time was allocated than in previous evaluations		18.0%	4
Same amount of time was allocated for this evaluation as for previous evaluations		40.0%	10
Less time was allocated for this evaluation as for previous evaluations		36.0%	9
Can't compare (I have not conducted any earlier evaluations for USAID)		8.0%	2
answered question			25
skipped question			0





12. How would you characterize the budget resources allocated for the most recent USAID evaluation for which you were the Team Leader?

		Response Percent	Response Count
More resources were allocated than in previous evaluations		18.0%	4
Same amount of resources were allocated for this evaluation as for previous evaluations		36.0%	9
Less resources were allocated for this evaluation as for previous evaluations		40.0%	10
Can't compare (I have not conducted any earlier evaluations for USAID)		8.0%	2
answered question			25
skipped question			0



13. In the SOW for the most recent USAID evaluation for which you were the Team Leader how were data collection methods addressed?

		Response Percent	Response Count
The SOW dictated the data collection methods to be used		16.0%	4
The SOW suggested data collection methods and asked us for comments or other ideas		60.0%	15
The SOW did not suggest data collection methods; it asked the team to suggest them		24.0%	6
The SOW did not address data collection at all		0.0%	0
answered question			25
skipped question			0





14. In the SOW for the most recent USAID evaluation for which you were the Team Leader how were data analysis methods addressed?

		Response Percent	Response Count
The SOW dictated the data analysis methods to be used		8.3%	2
The SOW suggested data analysis methods and asked us for comments or other ideas		33.3%	8
The SOW did not suggest data analysis methods; it asked the team to suggest them		54.2%	13
The SOW did not address data analysis at all.		4.2%	1
answered question			24
skipped question			1

15. Was the SOW for the most recent USAID evaluation for which you were the Team Leader structured around a list of QUESTIONS the team was expected to address?

		Response Percent	Response Count
Yes		96.0%	24
No		4.0%	1
answered question			25
skipped question			0

16. If the evaluation SOW was structured around a list of questions, how did it compare to earlier USAID evaluation SOWs from which you have worked

		Response Percent	Response Count
More questions		37.5%	9
Same number of questions, roughly		45.8%	11
Fewer questions		4.2%	1
Can't compare (either was your first evaluation for USAID or others were not structured around questions)		12.5%	3
answered question			24
skipped question			1

17. If the evaluation SOW was not structured around evaluation questions, which of the following was the primary focus for the evaluation as you understood it?

	Response Percent	Response Count
A list of issues the team was asked to address	0.0%	0
A list of objectives for the evaluation	100.0%	1
Other (please specify)		1
answered question		1
skipped question		24

18. For the most recent USAID evaluation for which you were the Team Leader, how would you characterize the amount of time your team had to prepare for data collection and analysis BEFORE heading for the field?

	Response Percent	Response Count
Virtually no time	20.0%	5
Slightly less than needed	48.0%	12
Adequate	28.0%	7
More than was needed	4.0%	1
answered question		25
skipped question		0



19. For the most recent USAID evaluation for which you were the Team Leader, were you asked to complete any of the following tasks before field work/data collection for the evaluation began?

	Yes	No	Rating Count
Analyze existing data, summarize what is already known, and identify what gaps remain that field work will be used to address	45.5% (10)	54.5% (12)	22
Present a detailed version of the evaluation design for USAID approval prior to field work	70.8% (17)	29.2% (7)	24
Sign a "No Conflict of Interest" form	60.9% (14)	39.1% (9)	23
		answered question	25
		skipped question	0

20. For the most recent USAID evaluation for which you were the Team Leader, how would you characterize the amount of time your team had for:

	Virtually no time	Slightly less than needed	Adequate	More than was needed	Rating Count
Data collection	4.5% (1)	40.9% (9)	54.5% (12)	0.0% (0)	22
Data analysis before completing and submitting a draft report	18.7% (4)	45.8% (11)	37.5% (9)	0.0% (0)	24
			answered question		25
			skipped question		0

21. For the most recent USAID evaluation for which you were the Team Leader, were you asked to provide a briefing to USAID or the firm that managed the evaluation on the preliminary findings, conclusions, and recommendations BEFORE writing the draft report?

		Response Percent	Response Count
Yes		92.0%	23
No		8.0%	2

answered question 25

skipped question 0

22. For the most recent USAID evaluation for which you were the Team Leader, how would you characterize USAID's review of your evaluation report from the following perspectives:

	Less thorough than in previous evaluations	About as thorough as in previous evaluations	More thorough than in previous evaluations	Rating Count
Structure (adherence to USAID's outline or template)	16.7% (4)	62.5% (15)	20.8% (5)	24
Strength of the evidence used when presenting findings	12.5% (3)	45.8% (11)	41.7% (10)	24
Adequacy of the linkages between evaluation findings, conclusions, and recommendations	8.3% (2)	58.3% (14)	33.3% (8)	24
Specificity and practicality of the evaluation recommendations	12.5% (3)	54.2% (13)	33.3% (8)	24

answered question 24

skipped question 1

23. Do you have any additional comments you would like to share in regards to your most recent experience as a USAID evaluation Team Leader compared to previous USAID evaluations you worked on? Please use the space provided below?

**Response
Count**

18

answered question 18

skipped question 7

Page 2, Q4. How many times have you served as a USAID evaluation Team Leader?

1	ten or more	May 25, 2013 10:26 AM
2	3	May 23, 2013 3:18 PM
3	3	May 23, 2013 10:47 AM
4	20	May 23, 2013 10:29 AM
5	3 - 4	May 22, 2013 4:04 PM
6	Five	May 22, 2013 3:11 PM
7	14	May 22, 2013 2:48 PM
8	4	May 22, 2013 2:08 PM
9	6	May 22, 2013 1:43 PM
10	2	May 22, 2013 12:24 PM
11	6	May 22, 2013 12:09 PM
12	7-8	May 18, 2013 2:34 PM
13	2	May 18, 2013 12:46 PM
14	At least 5 times	May 17, 2013 6:47 AM
15	two-three	May 16, 2013 11:23 PM
16	3	May 16, 2013 10:04 PM
17	Once	May 16, 2013 5:15 PM
18	more than 10	May 16, 2013 4:03 PM
19	About 5	May 16, 2013 3:50 PM
20	2	May 16, 2013 3:39 PM
21	once	May 16, 2013 3:23 PM
22	3	May 16, 2013 3:20 PM
23	twice	May 16, 2013 3:17 PM

Page 3, Q6. Please indicate the key factors that you feel have been most instrumental in your being selected to serve as a Team Leader for USAID evaluations?

1	My ability to deliver a high quality report on time	May 23, 2013 10:30 AM
2	Human Relations, communication, Cross-cultural experience, Peace Corps experience, personal contacts, breadth of experience with many different international agencies (UNICEF, UNESCO, UNDP, World Bank, Asian Development Bank, Global Partnership for Education, Fast Track Initiative)	May 18, 2013 2:34 PM
3	All the above have likely been important. My country experience (Afghanistan) may have been the biggest factor - along with team management (one team had four other people)	May 18, 2013 12:47 PM
4	writing and organizational skills	May 17, 2013 6:48 AM
5	proficiency in the language of the target community	May 16, 2013 3:25 PM
6	Area and language competence	May 16, 2013 3:20 PM

Page 6, Q9. Who provided you with the information you received on USAID evaluation quality standards?

1	Both USAID staff and general awareness of USAID procedures myself	May 25, 2013 10:28 AM
2	I'd been tracking USAID and MCC policies	May 22, 2013 2:51 PM
3	Both - but I cannot check both boxes	May 18, 2013 12:49 PM

Page 6, Q10. Which specific document or documents contained guidance on USAID's evaluation standards? (Check all documents you were given)

1	Standards of the firm that organized the evaluation team	May 16, 2013 2:06 PM
---	--	----------------------

Page 9, Q17. If the evaluation SOW was not structured around evaluation questions, which of the following was the primary focus for the evaluation as you understood it?

1	evaluation of how the implementing partner achieved the results as specified in their project SOW	May 25, 2013 10:30 AM
---	---	-----------------------

Page 13, Q23. Do you have any additional comments you would like to share in regards to your most recent experience as a USAID evaluation Team Leader compared to previous USAID evaluations you worked on? Please use the space provided below?

- | | | |
|----|---|-----------------------|
| 7 | One issue I have encountered recently is that USAID staff take issue with recommendations. I expect them to have questions and disagreements with findings as those are a product of interpretation. Recommendations, however, are from the evaluation team to USAID. They can accept or reject recommendations, or say they are unsubstantiated, but I am always surprised when they ask the team to remove a recommendation or change it. This is an areas which would benefit from more clarification for both USAID and evaluation teams. | May 22, 2013 1:55 PM |
| 8 | The SOW of the evaluations are being being better thought through. | May 22, 2013 12:28 PM |
| 9 | In the past 5 to 7 years, and more broadly in the last decade, the amount of time to implement USAID evaluations has in general radically lessened given the tasks to be done. There seems to be an increasing urgency to get evaluations over and out of the way. Evaluations seem to be more and more carried out simply to get them over with; to satisfy less-than-adequate time frames for implementation, and to satisfy increasing monetary constraints. Despite good guidance from consulting firms, who also seem increasingly "squeezed" by abrupt timing and planning changes over which they have little control, it seems increasingly evident to this consultant that financial constraints and fiscal control personnel (Contract Officers) with little knowledge of the evaluation issues have been controlling what is done. The quality of USAID evaluations suffers when consultants and firms are force-fitting too many activities and issues into increasingly inadequate time periods. Having for various reasons lost a lot of very experienced staff, USAID supervisory personnel have shown their own lack of knowledge, and most importantly wisdom, compared with a DECADE AGO. This is frankly sad. Finally, having worked for many other international agencies, it is clear that USAID (and other agencies as well) have "blinders on" and "tunnel vision" when it comes to using the experience.... and most importantly, wisdom, accumulated by similar agencies. How effectively and efficiently are "lessons learned" becoming "lessons implemented"? It could be better. Perhaps this survey will help. | May 18, 2013 2:38 PM |
| 10 | USAID staff were clearly very busy. To encourage review and critique we prepared an eight page booklet which summarized key findings and recommendations and this booklet was distributed by our USAID contacts to the larger audience. | May 18, 2013 12:55 PM |
| 11 | Recent evaluations I have worked on placed less emphasis on field site visits and more on other data sources (interviews, frequently by phone, surveys) | May 17, 2013 6:53 AM |
| 12 | The idea of random surveys of participants and getting them disclose sensitive business information was unrealistic. To do a good quantitative evaluation with surveys, you need a good baseline survey prior to starting the project. Even then, the results are not likely to fully access the successes or failures of the project. Often a somewhat wide ranging oral interview of contacts is necessary to find out the real story and to get leads on additional pertinent questions to explore. Running a formal survey in some countries makes business contacts suspicious (are you linked to tax collectors?) and sometimes doesn't elicit very useful data. This issue was similar in several evaluations I was involved in. The m&e designs could be overly complex, could sometimes waste time or lead to | May 16, 2013 10:19 PM |

Page 13, Q23. Do you have any additional comments you would like to share in regards to your most recent experience as a USAID evaluation Team Leader compared to previous USAID evaluations you worked on? Please use the space provided below?

activities not useful for the project (e.g. attempts to bloat training numbers) and would have been more useful if less complex. Certainly baselines and indicators are necessary but often they do not provide a true picture of successes and failures. They do provide a framework for more subjective review and guidance and revision throughout the project but generally do not provide a complete picture by any means. They are useful in summarizing results across projects. In sum, they are useful but their usefulness and effectiveness can be exaggerated.

- | | | |
|----|--|----------------------|
| 13 | My last evaluation was painful because the Mission claimed the project designed was imposed on them from AFR Bureau, all the senior USAID staff had already left the project, and the Mission Director disrupted the project schedule by 9 months due to fear of host-country sector corruption. | May 16, 2013 6:23 PM |
| 14 | time allocated to complete the evaluation keeps getting shorter | May 16, 2013 4:12 PM |
| 15 | Several people at USAID reacted unprofessionally to criticisms of the program we were evaluating. | May 16, 2013 3:53 PM |
| 16 | Usaid morocco took over two months to respond to the preliminary draft report and then nearly two months to respond to the revised report. This worked a serious heartship on the team leader and local consultants. Also it seemed that the contact q person at the mission (not a direct hire) was so closely connected to the project contrator that he was reluctant to accept critical evaluation and kept asking to tone it down | May 16, 2013 3:42 PM |
| 17 | My draft report was submitted to implementing agencies (CARE, World Relief). In the earlier evaluation, I never saw what report was sent by CARE to USAID. In the most recent one (for WR) I saw the final draft elaborated by WR, with th eoption for corrections. | May 16, 2013 3:24 PM |
| 18 | USAID and staff of USAID provided absolutely no comments or feedback on the draft evaluation report and final report. Not one word. | May 16, 2013 2:12 PM |

Annex E. Group Interview Transcript Summaries

Regional Bureaus Stakeholder Consultation Session

On May 2, 2013 MSI conducted a stakeholder consultation session with representatives of USAID's Regional Bureaus. The session was held in Washington DC at USAID's offices and was attended by five individuals representing the following bureaus: Africa; Latin America and the Caribbean; Eastern Europe and Eurasia; and the Office of Afghanistan and Pakistan. The session began with a short PowerPoint presentation introducing the evaluation study and an exercise where participants were asked to rank the five most important factors contributing to improvements in evaluation quality in their bureau. The factors were provided in a checklist format and were organized based on the evaluation process elements depicted in the Evaluation Process Wheel diagram introduced in the methods section of this report. This same diagram was also used to facilitate the discussion about evaluation quality at USAID, focused around the steps in the evaluation process.

The session began with the group members being asked if they had seen any changes at the beginning of the wheel, or the Initial Planning, SOW Development, and Team Composition stages. The discussants emphasized the importance of the quantity of questions being asked in evaluations and that Statements of Work (SOWs) with a high number of questions tends to lead to increased breadth of evaluation scope, not depth in the scope of evaluations. Participants also thought that a large number of questions may reflect a lack of planning, while fewer questions could represent more focused, relevant, and well-planned SOWs. There was a general consensus that the number of questions did not affect the cost of the evaluation, or the amount of time allocated to perform the evaluation. One group member stated that SOWs were improving following the peer-review process required by USAID Forward; while another said that this caused the numbers of questions to go down, but did not necessarily improve their quality. There was also a mention, however that a potential drawback of the peer-review process is that a review by someone who does not understand evaluations may have detrimental effects on SOWs such as potentially adding more questions. For this reason, there was agreement that there exists a definite need to train reviewers to provide consistent and high quality peer-reviews of SOWs. The peer review process, while no longer mandatory, was perceived as a collaborative relationship rather than an imposition from Washington, was well-received, has improved relationships, and continues to be implemented. Though this is a positive development, the varying degree of quality and consistency of the peer-review was also emphasized. Additionally, it was stated that the quality of the evaluation teams had generally remained the same over the last few years, and if composition had changed in the SOW, that was not resulting in higher quality or more appropriate team members as the same type of people are being hired to conduct evaluations as have been conducting them for years.

There was unanimous agreement that the methodology of SOWs had changed, though these changes were across the board with some Missions becoming more directive and some becoming less directive and more open in asking evaluators to come up with the methodology. While it was expressed that there was a desire for Missions to be more directive in terms of methodological rigor, it was also emphasized that if the Mission does not possess the capacity to develop a rigorous methodology, it is better to leave it to the evaluator rather than insert their own. This a positive change as a lack of capacity in evaluation design at the Mission level inhibits quality methodological directives. Group members all agreed that there was a noticeable increase in mixed methods, regardless of the Mission becoming more or less directive, and an improvement in the management process overall. One of the biggest improvements was a shift towards allowing more time for analysis, rather than going almost directly from data collection to writing. Analysis is now taking on a larger role and the evaluation timeline is improving.

The next subject discussed was the quality of draft and final reports. Interviewees agreed that they had not seen much change in this regard, even after the introduction of recommended structures. Good reports continued to be of good quality, and bad reports continued to be of poor quality, just in a newly styled

format. Participants claimed there is a disconnect between improvement in the quality of SOWs and changes in the quality of evaluation reports, and thought this was caused in part by a lack of capacity, in Washington and in the field, to successfully manage evaluations or recognize a quality evaluation report, as well as a lack of capacity among evaluators to conduct more robust evaluations. While there has been an effort to build the capacity of local evaluation firms, USAID often ends up doing most of the heavy lifting, which is very taxing on the time of Mission staff and has the potential to detract from the independence of an external evaluation. When asked if they felt evaluations by U.S. firms were of higher quality than local firms, respondents felt that this was not the case and the neither were producing quality reports. They felt that part of the problem is caused by hiring from the same group of evaluators as they have been for decades, both expatriates and locals, who lacked rigor before the new policy changes and continue to lack rigor after the changes. Thus, SOWs may have improved greatly, but the resulting report quality remains the same. A further reason for the disconnect between improved SOWs and evaluation reports is the desire by technical teams to have a technical expert lead an evaluation, rather than an evaluation expert. This can contribute to evaluations being led by professional field knowledge rather than by data. Another issue discussed by group members was the continued prevalence of short evaluations, lasting around two weeks, which do not allow for rigorous data collection and analysis; as well as under-funded evaluations.

Discussants were asked about any changes to evaluation dissemination and utilization towards the end of discussion. Most group members expressed that it was too early to see any changes in this regard, though they did note that there has been increased submissions to the DEC, though this is certainly not an indicator of improved report quality. The conversation ended with group members affirming that while the evaluation process has improved, an improvement in overall report quality has yet to follow.

Technical Bureaus Stakeholder Consultation Session

On May 2, 2013 MSI conducted a stakeholder consultation session with representatives of USAID's Technical Bureaus. The session was held at USAID headquarters and was attended by seven individuals representing the following bureaus: DCHA, E3, and Global Health. The session began with a short PowerPoint presentation introducing the evaluation study and an exercise where participants were asked to rank the five most important factors contributing to improvements in evaluation quality in their bureau. The factors were provided in a checklist organized around the evaluation process elements depicted in the Evaluation Process Wheel diagram introduced in the methods section of this report. This same diagram was also used to facilitate the discussion about evaluation quality at USAID.

The session began with the participants being asked if they had seen any changes in the Initial Planning, Statement of Work (SOW) Development, or Team Composition stages of the diagram. The discussion began by members stressing that with the new approach to the Program Cycle and the Evaluation Policy came a new and different approach to evaluation, causing projects to be shaped just as much by evaluations as evaluations are by projects. Discussants stressed that while the effect on overall quality may not be immediately evident, there is more opportunity for staff to engage in the conversation and increased support and outreach from Washington. This has resulted in a much more organized and thoughtful initiation to the evaluation process, including improved SOWs and team composition; however, past these initial stages the group members agreed that their confidence fell dramatically regarding any increase in quality.

Despite agreement among interviewees that there were improvements, or at least changes, in the early stages of evaluation, the conversation then moved towards some of the problems which they still see. Some of the issues that were brought up were: a failure to distinguish between Impact Evaluations and Performance Evaluations, Missions lacking personnel with evaluation expertise, the quality and quantity

of questions not improving, and a lack of specificity in SOWs. One participant posited that this could be because of their perception that an Implementation Plan of the Evaluation Policy was not rolled out; that they are not supporting their staff sufficiently; and that there is an overall lack of evaluation experts. For example, a Mission may now have a SOW peer-reviewed, but if that reviewer is not an evaluation specialist then the quality may not improve and in fact could be detrimental. Peer-reviewing itself is insufficient without more training and direction. One particular issue stressed by interviewees was the difficulty in recruiting a team with the right combination of experience and education. While the situation is better in global health, where many personnel have had evaluation training required as part of their background and education, in other sectors it is difficult to find people who possess evaluation as well as sector specific expertise. This makes it difficult do much past field visits and surveys.

When the conversation moved to data collection and analysis, group members noted that the changes are not as dramatic as those in the initial stages. Technology was mentioned as changing data collection for the better, however, this remains the exception rather than the rule. Discussants agreed that, while the situation has improved somewhat with regards to data quality, it remains difficult and expensive to implement a high quality impact evaluation, with one participant mentioning that people are often very unaware of the additional costs associated with Impact Evaluations. There has not been an increase in the variety or quality of data collection methods used, with many implementers sticking with what they are already familiar with. Evaluations remain mostly anecdotal rather than emphasizing empirical data and methodological rigor.

With regards to draft and final reports, group members expressed that is too soon to see changes based on new Evaluation Policy, which has so far mostly affected evaluation design and planning. As was mentioned previously, the quality of the final report cannot be expected to improve without improving the quality of the inputs, namely individuals with high levels of training and experience in evaluation. It was emphasized that there is no shortage of personnel who are comfortable with monitoring, but when it comes to the “E” part of M&E, there remains a lack of people able to really boost the quality of evaluations. When asked about any changes in evaluation utilization, one discussant mentioned that her bureau chief is really pushing utilization through brown bags and monthly journal club, though this focuses on the broader literature and not USAID evaluations specifically. It was also mentioned that CDCS has emphasized the use of evaluations in its strategy process, though at an Agency-wide level the utilization has not significantly changed.

Development Firms Stakeholder Consultation Session – Interview #1

On May 7, 2013 MSI conducted a stakeholder consultation session with representatives from International Development Firms. The session was held in Washington, DC at MSI’s offices and was attended by twelve individuals representing the following companies and organizations: JBS International, Catholic Relief Services (CRS), Development Alternatives International (DAI), FHI360, Creative Associates, Volunteers for Economic Growth Alliance (VEGA), Abt Associates, and Management systems International (MSI). The session began with a short PowerPoint presentation introducing the evaluation study and asked to consider the different factors contributing to improvements in evaluation quality. The factors were provided in a checklist format and were organized based on the evaluation process elements depicted in the Evaluation Process Wheel diagram introduced in the methods section of this report. This same diagram was also used to facilitate the discussion about evaluation quality at USAID.

The session began with the group members being asked if they had seen any changes at the beginning of the wheel, or the Initial Planning, SOW Development, and Team Composition stages. The quantity and quality of the questions posed in the SOW was the first topic broached by the discussants, who expressed

that while USAID may have reduced the number of questions that are included in an SOW, those questions are loaded with multiple component parts that need to be deconstructed. This also can happen in Results Frameworks, with multiple development objectives being included as one; these loaded development objectives are then often not considered when drafting SOWs, even though they include evaluation questions embedded in them. Another concern raised about SOWs was that sometimes the questions asked do not align with what USAID wants to be evaluated; this is potentially caused by a lack of reference to the Theory of Change and can lead to evaluations going down tangential paths.

One concern that was raised early on in the conversation was the feeling that USAID Missions are citing a need for Impact Evaluations, when that may not be the best design for the project. Oftentimes development projects and Impact Evaluations do not marry well due to the need to withhold interventions from certain groups, something that often is counterintuitive to development goals. There seems to be a disconnect between how the language and evaluation terminology is understood by the evaluation implementers and by the authors of SOWs, which can cause confusion. Another issue raised regarding Impact Evaluations was that while USAID understands the need for a quality Impact Evaluation to be started at the same time as the project, this is a difficult task in reality due to procurement issues, project implementers not agreeing to the conditions necessary to implement a randomized control trial, difficulty creating budgets and designs due to vague RFPs, and a lack of Impact Evaluations being built into the initial project design. It was also expressed by discussants that there seems to be confusion within USAID as to when to conduct a Performance Evaluation and when to do an Impact Evaluation, and that Impact Evaluations were being driven by a notion of compliance and not necessarily because they would fit well with the project design. Another manifestation of this notion was the impression that those authoring RFPs want to see impacts without grasping the full methodological and budgetary ramifications of Impact Evaluations. However, these struggles are indicative that USAID is taking evaluations to a new level and that the process is promising to improve in coming the coming years.

The conversation turned next to SOWs, with agreement that Missions and project officers are thinking more about the questions they want answered with an evaluation, a key point in the initial planning and SOW development and resulting in SOWs that are more weighted with analysis. Participants felt that improvements in SOW quality were higher in RFPs for third party evaluations than in the SOWs where the M&E plan is built into the project. It was reported that most SOWs are hybrids, or containing elements of both Performance and Impact Evaluations, asking for counterfactuals and the setting up of control groups as well as performance questions. One discussant proposed that these cases are caused by a misunderstanding of USAID guidance and the emphasis placed on evaluation language, not because they were intended to be hybrid evaluations. Two other concerns mentioned regarding SOWs were that they sometimes include cost analysis or a cost benefit analysis, which is seen as out of context with a performance evaluation, and that the persons authoring SOWs are not evaluation specialists or are very junior officers, sometimes resulting in muddled SOWs and confusion regarding the evaluation methods and goals. Because the discussion is at an early stage, the group felt that if the support of USAID remains there to take the practice forward that the quality of evaluation SOWs will continue to improve.

Evaluation Team composition was the next topic of conversation, and began with agreement that there was an increase in requests for evaluation specialists. There also started to be requests for Human Institutional Capacity Development (HICD) specialists and certifications. Apart from HICD, however, discussion participants expressed that they felt that the descriptions of what constitutes an evaluation specialist remained vague, and consist of varying combinations of number of years of experience and sector knowledge.

Continuing around the Evaluation Process Wheel, the time and cost of evaluations was discussed next. While many participants felt that the costs of evaluations had largely stayed the same, they did believe that the duration and timelines of evaluations had improved somewhat, with two to three weeks being added to the period of performance of evaluations. At this time many of the participants stated that in the

last two years there was buzz within their organizations and companies and interest by internal leadership around evaluations and improving their M&E plans and implementations.

The last main topic of conversation was pre-field work and methodologies in practice. Participants stated that while sometimes SOW authors will go back to the PMP and see what data is being collected to determine what is possible through an evaluation, at other times there is a lack of acknowledgment that baseline data does not exist causing evaluators to go back and collect baseline data after the fact, which can prove to be very difficult if not impossible. With regards to methodology itself, some discussants felt that while there has certainly been an increase in discourse, there remains hegemony in methodological practices with the randomized control trial considered the gold standard and little open-mindedness to other methods. One participant stated that in the last three to four evaluations conducted by the firm a multiplicity of methods were used including Survey Monkey, structured observations, and interviews with village leaders and teachers; while another participant stated that the methods used by the firm had not changed over the years. It was also stated that more project-level M&E personnel are familiar with SPSS software, and that the PMPs and systematic data collection is on the rise. One concern mentioned was that third party evaluators will come in and do a statistically insignificant number of interviews, for example, and ignore the rich amount of data that has already been collected by the implementing organization. Similarly, one firm reported that when implementing a project, from the beginning they prepare for an external evaluation. This makes them less vulnerable to the quality of an external evaluation and enables them to respond should they receive a poor evaluation report. Lastly, in terms of submitting draft and final evaluation reports, many participants agreed that there was an increase in back and forth with USAID which did not happen in the past and demonstrates that there is interest and that the reports are being read. Additionally, there is a sense of the increased desire on the part of USAID to translate evaluation findings into differences and improving results.

Development Firms Stakeholder Consultation Session – Interview #2

On May 8, 2013 MSI conducted a stakeholder consultation session with representatives from international development firms. The session was held in Washington, DC at MSI's offices and was attended by thirteen individuals representing the following companies and organizations: Checchi Consulting, Management Systems for Health (MSH), Education Development Center (EDC), The QED Group, Mitchell International, Development & Training Services Inc. (dTS), Development Alternatives, Inc. (DAI), and International Business & Technical Consultants Inc. (IBTCI). The session began with a short PowerPoint presentation introducing the evaluation study, and asked participants to consider the different factors contributing to improvements or changes in evaluation quality. The factors were organized based on the evaluation process elements depicted in the Evaluation Process Wheel diagram introduced in the methods section of this report. This same diagram was also used to facilitate the discussion about evaluation quality at USAID.

Starting with the initial planning stage, participants stated that some Missions have started to plan Impact Evaluations together with projects, and noted several examples of requests for proposals (RFPs) for projects and corresponding evaluations being released at the same time. Though initial planning varies from Mission to Mission, there was consensus that USAID has increased its efforts to get projects and their external evaluations on the same schedule. Related to this is the incorporation of baselines into the project startup phase, which according to discussants had not been done previously. One evolving aspect of the initial planning stage that several discussants were excited about was the introduction of a participatory design phase, already being implemented by some Missions, which gets all stakeholders involved in the evaluation design; this is called a Participatory Impact Pathways Analysis, or PIPA. Another positive change in the initial planning stage is the monitoring and inclusion of existing project

data in the evaluation design, as evaluations used to be designed without taking this preexisting data into account.

When asked about any changes in evaluation scopes of work (SOWs), participants noted many improvements such as SOWs being more clear, thoughtful, and well-written than they had been in the past. Specifically, SOWs now identify whether the assignment is an evaluation or an assessment, whereas this distinction was not always evident previously. This is reflective of the USAID Evaluation Course teaching participants how to write quality SOWs. Discussants stated that while they felt the Missions put more thought into what they want from evaluations when drafting SOWs, they do often draw too heavily on policy. It was expressed that the SOW schedules were ambitious and at times unrealistic, because of time it takes to planning and also contracting an evaluation team. With regards to Impact Evaluations, SOWs are now asking for attribution and include phrasing such as “rigorous evaluation”.

The conversation then turned to the cost and duration of evaluations and the composition of evaluation teams. It was expressed that while SOWs are improving in both rigor and evaluation methodology, there is still a lack of clarity on the actual costs of conducting a quality Impact Evaluation, demonstrated by unrealistically low budgets. This results in a lot of back and forth between USAID and the evaluator on the budget and subsequently delays the evaluation or reduces the rigorousness of the SOW. Though there was a general consensus that it is difficult to perform what USAID is requesting with the amount of money allocated, it was also agreed that evaluation timelines are improving. An increase in flexibility from USAID on timelines and team composition was appreciated, as one discussant stated that the firm would rather have more time and fewer team members. One challenge regarding timelines that was shared was an underestimation by USAID of the amount of time it would take USAID employees to come back with comments on draft evaluation reports and other internal processes. One discussant shared the view that the most important change in evaluation policy in recent years is team composition, with SOWs now requiring that team members have past evaluation experience. Though Team Leaders are now required to have some evaluation experience, concern was expressed that the Evaluation Specialist did not lead the team, as they may be better suited for designing the evaluation. The definitions of roles on the team and the qualifications for those roles have also improved dramatically, with only some broad language remaining. Qualifying descriptors were especially noted in the health sector. One concern raised was that SOWs often include qualification requirements that are nearly impossible to fulfill. The example was given of requiring 15 years of experience in a field that has not yet been around for 15 years. This has caused firms to submit personnel that do not fit the qualifications required and though the evaluation goes on, it leaves the impression that USAID is not certain of what it wants in an evaluation team.

There were several changes noted with regards to project timelines and methodologies. One positive change is that there are often two to three days now allotted for document review prior to deployment in the field; though this is often undercut by the fact that teams are sometimes not provided with these documents until they are out in the field. Some firms are beginning to organize early planning and team-building exercises, especially when team members are from various locales, and are including these meetings in their proposals. One discussant stated that after reviewing more than 50 evaluation SOWs, only about three of them required the team to present a final work plan and a presentation on the evaluation before being allowed to conduct the evaluation, and a similar number required that all work be completed in-country. Another participant stated that in one evaluation IQC, the firm is required to submit an inception report, demonstrating a clear plan of how to move forward, with every Task Order. Additionally, it was shared that some Missions are requesting reports on the evaluation’s findings to date, as well as RFPs that require the submission of sample instruments. However, participants felt that because developing quality evaluation instruments is very time consuming, a rapid proposal timeframe is not amenable with developing high-quality tools. Additionally, this requires a deal of costly work on the part of the evaluator before the contract has been awarded.

As mentioned previously, some firms feel that Missions draw too heavily on policy for identifying preferred evaluation methodologies, resulting in methods that may not be well-suited for the project at hand. The evaluation then becomes driven by an adherence to policy more than by the design of the project it is evaluating. Another example of this is Missions requesting quantitative data to be within a level of confidence in a performance evaluation, suggesting too heavy of a reliance on policy. There was consensus around the feeling that Missions were being pushed by policy to issue a certain number of evaluations per year, including one Impact Evaluation. However, participants reported that Missions are flexible and open to evaluators coming back to them with suggested changes to the methodology, reflecting a desire to learn and improve. It was also stated that generally, USAID is requesting more focus group discussions, key informant interviews, and surveys.

When submitting draft evaluation reports to USAID, participants noted that there is substantially more rigor in terms of matching evaluation findings to conclusions, and ensuring that the report presents actual findings and not just conclusions. Subsequently, the link between conclusions and recommendations is also examined. Though this systematic review is an improvement and is beneficial overall, some discussants felt that it could be done to a fault and has become unnecessarily formulaic, citing a heavy reliance on “checklists”. In certain cases, this resulted in the evaluator using a similar checklist as an outline tool for drafting the report. Several discussants stated their belief that USAID is on a learning curve with regards to evaluations and that the process and overall quality of evaluations will continue to improve.

Annex F. Team Composition

The following provides a brief description of each of the members of the MSI meta-evaluation research and coding team. MSI has on file a signed No-Conflict of Interest form for each of the team members that completed the study.

Molly Hageboeck (Team Leader and Technical Director) has significant experience working both for and with USAID on evaluations and evaluation quality and has conducted numerous evaluations including several meta-evaluations and similar evaluation quality review exercises including many of the reports referenced earlier in the design document. Ms. Hageboeck has served as Team Leader for previous meta-evaluations for USAID, including A Review of the Quality and Coverage of A.I.D. Evaluations, FY 1989 and FY 1990; Trends in International Development Evaluation Theory, Policy and Practices; and From Aid to Trade – Delivering Results.

Micah Frumkin (Team Manager and Senior Coder) has been working with USAID as an external consultant for more than five years and has had a focused attention on evaluations at USAID including as a team member on numerous studies reviewing the quality of evaluation reports and statements of work such as for the studies Quality Review of Recent USAID Evaluation Statements of Work, and Trends in International Development Evaluation Theory, Policy and Practices report, among others. Mr. Frumkin has been the facilitator for MSI's Certificate Program in Evaluation for several years and has assisted in the certification of more than 200 evaluators, the vast majority of which were USAID staff. He has contributed to numerous evaluations and co-wrote many of the documents in the USAID TIPS series on Monitoring and Evaluation including some of the new Technical Notes currently in production.

Stephanie Monschein (Quantitative Data Analyst and Coder) has been with MSI for over five years, where she has gained experience as an evaluation rater and SPSS data analyst for a variety of projects including From Aid to Trade – Delivering Results, MSI's agency-wide trade capacity building evaluation, in which substantive and M&E characteristics of over 200 projects were rated and examined, using a chi square test to identify differences from expected project and M&E characteristics. She has also designed and managed the M&E system and completed the Mid-Term evaluation for an MSI project in Zambia; served as the M&E Technical Manager for MSI's USAID/Kenya long term M&E Support Project; and as an M&E specialist, designing and developing content for an online and interactive website of M&E Tools, templates, and guidance for E3.

Adam Peterson (Coder) has been with MSI for nearly five years where he has managed various field programs, evaluations, and assessments as well as assisted in the design of evaluations, including survey design, sampling, data collection and analysis. Mr. Peterson has also studied evaluation methods in graduate classwork at Georgetown University.

Elizabeth Freudenberger (Coder) has been with MSI for nearly four years where she has managed numerous evaluations and used USAID evaluation report checklists to ensure MSI compliance with USAID standards. Ms. Freudenberger also has previous evaluation rating experience on MSI's Trends in International Development Evaluation Theory, Policy and Practices study for the USAID Evaluation Office.

Gwynne Zodrow (Coder) is part of MSI's Strategic Management and Performance Improvement practice area where she currently works to provide results-based management technical assistance to FDA and oversees the implementation of USAID's Africa Lead's M&E systems to collect, manage, and analyze data for reporting progress towards the objectives. Previously Ms. Zodrow has helped develop and conduct evaluations on food security and livelihood projects in Africa.

Ingrid Orvedal (Coder) has served as the Assessments, Monitoring, and Evaluation Advisor to USAID as part of the SUPPORT Project in South Sudan, implemented by MSI. As such, Ms. Orvedal managed over 25 evaluations and assessments of USAID-funded activities from 2010 to 2012, including SOW and methodology development and report writing and editing. In that role she became an experienced evaluation rater as MSI/South Sudan used evaluation checklists to score all evaluations and provide feedback to teams.

Jeremy Gans (Coder) has been with MSI for nearly five years. He has managed numerous evaluations and used USAID evaluation report checklists to ensure MSI compliance with USAID standards. Mr. Gans has previous evaluation rating experience on MSI's Trends in International Development Evaluation Theory, Policy and Practices study for the USAID Evaluation Office and Quality Review of Recent USAID Evaluation Statements of Work.

Leah Sly (Coder) is an International Recruiter at MSI, where she specializes in recruiting M&E specialists and staffing USAID evaluation projects. As a writer/editor, she has drafted final project reports, tracked project performance measures, and edited performance evaluation reports and annexes. For her Master of Public Policy degree she completed coursework in advanced statistical methods and analytical policy frameworks.

Mary Beth Allen-Yarbrough (Coder) has more than ten years working with USAID, including in PPC/Evaluation, both as a foreign and civil service officer with a focus on USAID evaluation policy implementation, Logical Frameworks and evaluation quality. Ms. Allen-Yarbrough is the co-author of USAID's meta-analysis Private Sector: Ideas and Opportunities : A Review of Basic Concepts and Selected Experience.

Paul Diegert (Coder) is a Project Manager at MSI with experience in monitoring and evaluation. Most recently he worked on a team of evaluators to conduct a meta-analysis of international NGOs' performance on unrestricted grants. Mr. Diegert has also helped develop a performance monitoring plan (PMP) for an education project in Mali, which used a quasi-experimental student achievement test as the top-level impact indicator.

Sarah Fuller (Coder) is a Project Manager at Management Systems International where she currently supports a monitoring and evaluation program. Her contributions to the numerous evaluation reports produced by the program have made her very knowledgeable of the content and structure of quality reporting as well as USAID evaluation expectations.

Annex G. Bibliography

- Blue, Richard; Cynthia Clapp-Wincek and Holly Benner. 2009. *Beyond Success Stories: Monitoring & Evaluation for Foreign Assistance Results*. Report prepared for USAID. Washington, DC. <http://pdf.usaid.gov/pdf_docs/PCAAB890.pdf>.
- Clapp-Wincek, Cynthia and Richard Blue. 2001. *Evaluation of Recent USAID Evaluation Experience*. Report prepared for USAID. Washington, DC. <<http://portals.wi.wur.nl/files/docs/ppme/PNACG632.pdf>>. Referred to as “Clapp-Wincek & Blue (1998 & 1999)”.
- Bollen, Ken. 2005. "Assessing International Evaluations: An Example From USAID's Democracy and Governance Program." *American Journal of Evaluation* 26.2: 189-203. Print. Referred to as “Bollen, 2006”.
- Eriksson, John and Krishna Kumar. 2011. *A Meta-Evaluation of Foreign Assistance Evaluations*. Prepared for the Office of the Director of U.S. Foreign Assistance, Department of State. Washington, DC. http://pdf.usaid.gov/pdf_docs/PCAAC273.pdf. Referred to as “Kumar and Eriksson (2009)”.
- Frumkin, Micah, Emily Kearney, and Molly Hageboeck. 2010. *Quality Review of Recent USAID Evaluation Statements of Work*. Prepared for USAID. Washington, DC: Management Systems International. <http://pdf.usaid.gov/pdf_docs/PNADX372.pdf>.
- Greene, Katrina. 1999. *Narrative Summary of FY97 Evaluations*. Report prepared for USAID/R&R. Washington, DC. No URL or hard copy was found. Referred to as “Greene (1993 – 1997)”.
- Hageboeck, Molly. 1992. *A Review of the Quality and Coverage of AID Evaluations, FY 1989 and FY 1990*. Report prepared for USAID. Washington, DC: Management Systems International. Print. Referred to as “MSI (1989 & 1990)”.
- Hageboeck, Molly. 2010. *From Aid to Trade: Delivering Results, A Cross-Country Evaluation of USAID Trade Capacity Building*. Report Prepared for USAID. Washington, DC: Management Systems International. <http://pdf.usaid.gov/pdf_docs/PDACR202.pdf>. Referred to as “MSI, 2010”.
- Hageboeck, Molly. 2009. *Trends in International Development Evaluation Theory, Policy and Practices*. Report prepared for USAID. Washington, DC: Management Systems International. <http://pdf.usaid.gov/pdf_docs/PNADQ464.pdf>. Referred to as “MSI (2005 to 2008)”.
- Hopstock, Paul J., Allan C. Kellum, and Malcolm B. Young. 1989. *Review of the Quality of AID Evaluations, FY 1987 and FY 1988*. Report prepared for USAID. Washington, DC: Development Associates. <http://pdf.usaid.gov/pdf_docs/PNABC321.pdf>. Referred to as “Development Associates (1987 & 1988)”.

Instrument used in MSI's Certificate Program in Evaluation, updated in April 2006

Instrument used in MSI's Certificate Program in Evaluation, updated in November 2007

Instrument used in MSI's Certificate Program in Evaluation, updated in August 2009

Instrument used in MSI's Certificate Program in Evaluation in Juba, South Sudan in June 2011

Instrument currently used in USAID's EES and EPM course and available through the USAID website

Internal MSI document containing data from 2008 evaluations that were not included in the *Trends in International Development Evaluation, Theory, Policy and Practices* report. Referred to as "MSI (2008, remaining months)".

Kerley, Janet. 2000. *A Report on the Use of Evaluations in the Europe and Eurasia Bureau/USAID; An Inventory of Evaluations*. Report prepared for the Office of Program Coordination and Strategy, Europe and Eurasia Bureau, USAID. Washington, DC. Referred to as "Kerley (1999 – 2004)".

Savedove, William D., Ruth Levine, and Nancy Birdsall. 2006. *When Will We Ever Learn? Improving the Lives through Impact Evaluation*. Washington, DC: The Center for Global Development.
<http://www.cgdev.org/sites/default/files/7973_file_WillWeEverLearn.pdf>.

Triton. 1982. *Final Report: Development of a Quality/Completeness Scoring Instrument for USAID Evaluations*. Report prepared for USAID. Washington, DC: Triton.
<http://pdf.usaid.gov/pdf_docs/PNABD049.pdf>.

Triton. 1984. *Final Report: Findings Compendium and Analysis of FY82 AID Evaluation Reports*. Report prepared for USAID. Washington, DC: Triton. Print. Referred to as "Triton (1982)".

Triton. 1985. *Final Report: Analysis of the Distribution of Quality/Completeness Scores of FY83 AID Evaluation Reports*. Report prepared for USAID. Washington, DC: Triton. Print. Referred to as "Triton (1983)".

USAID. 2003. *ADS Chapter 203 Assessing and Learning*. Policy, Planning, and Learning. Washington, DC. Partial revision date: January 21, 2003.

USAID. 2008. *ADS Chapter 203 Assessing and Learning*. Policy, Planning, and Learning. Washington, DC. Partial revision date: September 1, 2008.

USAID. 2010. *ADS Chapter 203 Assessing and Learning*. Policy, Planning, and Learning. Washington, DC. Partial revision date: April 2, 2010.

USAID. 2012. *ADS Chapter 203 Assessing and Learning*. Policy, Planning, and Learning. Washington, DC. Partial revision date: November 2, 2012.

Yin, Robert K.; and Carol H. Weiss. 1988. *Preliminary Study of the Utilization of AID's Evaluation Reports*. Report prepared for USAID. Washington, DC.
<http://dec.usaid.gov/index.cfm?p=search.getCitation&CFID=9801169&CFTOKEN=36491238&id=s_DD19F9D6-D566-FC5C-D98314451C9F0EBB&rec_no=52769>.