



EVALUATION

Examination of Foreign Assistance Evaluation Efforts at the Department of State: A Combined Performance and Meta-Evaluation

April 2018

This publication was produced at the request of the United States Department of State. It was prepared independently by ME&A.

EXAMINATION OF FOREIGN ASSISTANCE EVALUATION EFFORTS AT THE DEPARTMENT OF STATE

A Combined Performance and Meta-Evaluation

April 6, 2018

IDIQ Contract Number: D17PC00465

Technical and Advisory Services for Program Evaluation Requirements

Task Order Number: D17PD01299

Submitted by:

Dr. Gary Woller, Team Leader

Dr. Lawrence Robertson

Christopher Coffman

Dereje Gebreegziabher

Andres Rueda

DISCLAIMER

The authors' views expressed in this publication do not necessarily reflect the views of the United States Department of State or the United States Government.

CONTENTS

Executive Summary.....	v
Evaluation Purpose and Evaluation Questions	v
Background	v
Evaluation Design, Methods, and Limitations	v
Findings and Conclusions.....	vi
Recommendations	x
1.0 Evaluation Purpose and Evaluation Questions	1
1.1 Evaluation Purpose and Audiences.....	1
1.2 Evaluation Questions	1
2.0 Background	2
3.0 Evaluation Design, Methods, and Limitations	2
3.1 Evaluation Design and Methods	2
3.1.1 Inception Phase and Document Review	3
3.1.2 Meta-Evaluation Phase.....	3
3.1.3 Performance Evaluation Phase	5
3.1.4 Data Analysis Phase	6
3.1.5 Report Writing Phase.....	6
3.2 Evaluation Team	6
3.3 Methodological Limitations	7
3.3.1 Actions Taken to Mitigate Methodological Limitations.....	8
4.0 Findings and Conclusions.....	9
4.1 EQ 1: What does the structure and content of evaluation reports tell us about the strengths and weaknesses of methodology, design, and analysis in evaluations?	9
4.1.1 Overall Evaluation Report Quality	9
4.1.2 Evaluation Report Quality by Quality Category	11
4.1.3 Evaluation Report Quality by PQI	12
4.1.4 Evaluation Report Quality by Report Characteristic.....	15
4.2 EQ 2: To what degree are foreign assistance evaluation findings and recommendations being used by bureaus, and what factors contribute to their use or non-use?	16
4.2.1 Perceived Importance of Evaluating Bureau Programs	16
4.2.2 How Bureaus Have Used Evaluations	17
4.2.3 Factors that Hinder Bureau Evaluation Use.....	20
4.3 EQ 3: To what extent have bureaus met the baseline requirements of the policy for planned and completed evaluations and staffing and utilizing the Bureau Evaluation Coordinator position?	25
4.3.1 Familiarity with and Support for DoS Evaluation Policy	26
4.3.2 Bureau Compliance with DoS Evaluation Policy Requirements	27
4.3.3 Use of BECs and BEC Roles, Responsibilities, and Visibility	30
4.3.4 BEC and Bureau-Wide Evaluation Capacity	32
4.3.5 Capacity and Evaluation Implementation Challenges	34
4.3.6 Senior Leadership Buy-In	35
4.3.7 Change in Evaluation Practice Since 2012	37
4.4 EQ 4: To what degree have bureaus used F’s capacity building tools, classes, papers, and other technical assistance to increase evaluation capacity and improve evaluation? ...	39

4.4.1	Familiarity with F Evaluation Tools	39
4.4.2	Use and Usefulness of F Evaluation Tools	39
4.4.3	Use of F Consultations/Technical Assistance and Participation in the Evaluation Community of Practice	41
4.4.4	Usefulness of F Consultations/Technical Assistance and the Evaluation Community of Practice	42
4.4.5	Usefulness of F and BP Evaluation Training Courses	45
4.4.6	Recommendations for Improving F Evaluation Tools and Evaluation CoP.....	45
4.5	EQ 5: Looking forward to implementation of FATAA, what areas are likely to require more attention or focus from F?	46
4.5.1	Areas Likely to Require More Support or Attention From F.....	46
5.0.	Recommendations	49
5.1	Evaluation Question 1.....	49
5.2	Evaluation Question 2.....	50
5.3	Evaluation Question 3.....	51
5.4	Evaluation Question 4.....	51
5.5	Evaluation Question 5.....	52
Annexes	53
	Annex I: Evaluation Statement of Work	54
	Annex II: Bibliography	60
	Annex III: Evaluation Matrix.....	63
	Annex IV: Meta-Evaluation Checklist.....	65
	Annex V: Online Survey Questionnaire.....	83
	Annex VI: KII and FGD Discussion Guides	108
	Annex VII: List of Persons Participating in KIIs and FGDs.....	113
	Annex VIII: Evaluation Methodology	117
	Annex IX: Online Survey Results	125
	Annex X: Relationships between Survey Responses and Different Respondent Categories.....	135
	Annex XI: Disclosure of any Conflicts of Interest	138

LIST OF TABLES

Table 1.	Distribution of Evaluation Reports Reviewed for the Meta-Evaluation by Year of Completion.....	3
Table 2.	Number of Principal and Supplemental Checklist Indicators by Scoring Category.....	4
Table 3.	Percentage of Reports Rated as Good, Fair, and Poor for Each of the Nine Checklist Categories.....	11
Table 4.	Ranking and Categorization of Checklist Principal Quality Indicators	12
Table 5.	Factors Hindering Bureau Evaluation Use (Unprompted).....	21
Table 6.	Compliance with Evaluation Policy Requirement to Conduct at Least One Evaluation per Year	27

LIST OF FIGURES

Figure 1.	Distribution of Overall Meta-Evaluation Scores	10
Figure 2.	Number of Good, Fair, and Poor Evaluation Reports	10
Figure 3.	Extent to Which the Bureaus Have Used Evaluations for Different Purposes (N=60).....	18
Figure 4.	Factors Hindering Bureau Evaluation Use (Prompted).....	20
Figure 5.	Leadership Support for the DoS Evaluation Policy (N=58).....	26
Figure 6.	Extent to which the Evaluation Policy Applies to the Bureau (N=58).....	27
Figure 7.	Extent to Which the Bureau Has Complied With the Evaluation Policy (N=58)	28

Figure 8. Use of and Usefulness of BEC Consultations	31
Figure 9. Familiarity with F Evaluation Tools (N=46)	39
Figure 10. F Evaluation Tools and Their Perceived Usefulness.....	40
Figure 11. Use of Different F Evaluation Resources.....	42
Figure 12. Frequency of Participation in the Evaluation Community of Practice (N=45).....	42
Figure 13. Frequency and Form of Participation in the Evaluation Community of Practice (N=45)	42
Figure 14. Usefulness of F Consultations and Participation in the Evaluation Community of Practice	43

ACRONYMS

Acronym	Description
ANOVA	Analysis of Variance
BEC	Bureau Evaluation Coordinator
BP	Bureau of Budget and Planning
CoP	Community of Practice
DAS	Deputy Assistant Secretary
DE	Diplomatic Engagement
DoS	U.S. Department of State
EQ	Evaluation Question
ET	Evaluation Team
F	Office of U.S. Foreign Assistance Resources
FA	Foreign Assistance
FAM	Foreign Affairs Manual
FATAA	Foreign Aid Transparency and Accountability Act
FGD	Focus Group Discussion
FSI	Foreign Service Institute
GAO	Government Accountability Office
IDIQ	Indefinite Delivery/Indefinite Quantity
IGCE	Independent Government Cost Estimate
KII	Key Informant Interview
MfR	Managing for Results
M&E	Monitoring and Evaluation
MOPAN	Multilateral Organization Performance Assessment Network
OIG	Office of Inspector General
PEPFAR	The President's Emergency Plan for AIDS Relief
PDAS	Principal Deputy Assistant Secretary
PQI	Principal Quality Indicator
SOW	Statement of Work
SQI	Supplemental Quality Indicator
Sub-EQ	Sub-Evaluation Question
TA	Technical Assistance
U.S.	United States

EXECUTIVE SUMMARY

Evaluation Purpose and Evaluation Questions

The United States Department of State (DoS) Office of U.S. Foreign Assistance Resources (F) contracted ME&A to conduct a combined performance and meta-evaluation. The purpose was to examine the evaluation efforts since 2012 of the 26 DoS bureaus that use Foreign Assistance (FA) funds. The meta-evaluation assessed the quality of completed evaluations through an analysis of their strengths and weaknesses on the basis of evaluation report content. The performance evaluation assessed: 1) the extent to which evaluation findings and recommendations are used and what obstacles, if any, hinder their use; 2) the extent to which the Department's efforts to institutionalize evaluation have been effective in building evaluation expertise and capacity in bureaus; and 3) where F should direct its assistance to bureaus moving forward, based on the strengths and weaknesses of bureau evaluation practices. The performance and meta-evaluations addressed the five evaluation questions (EQs) found below, under Findings and Conclusions.

Background

F established the first department-wide Evaluation Policy in 2012. F and the Bureau of Budget and Planning (BP) then developed a revised Evaluation Policy in 2015 and updated guidance and support for the evaluation. In November 2017, in conjunction with BP, F revised the Evaluation Policy to respond to the requirements of Foreign Aid Transparency and Accountability Act (FATAA), by placing evaluation as part of a continuum for planning, managing, monitoring, evaluating, and learning.

Since 2013, F and BP have offered evaluation courses for the Department and provided evaluation technical assistance to bureaus through consultations. Per the Evaluation Policy, bureaus and independent offices have assigned bureau evaluation coordinators (BECs), who support the planning and management of evaluations in their bureaus and offices. F and BP further manage an evaluation community of practice (CoP) for the Department, which meets monthly to discuss issues and share information. F also manages online resources to support bureaus with additional information on evaluation.

As a result of these developments and requirements of the Evaluation Policy, bureaus and independent offices conducted more than 180 evaluations from 2012 through 2017.

Evaluation Design, Methods, and Limitations

To conduct the performance and meta-evaluation, ME&A employed a mixed-methods design. For the meta-evaluation, the evaluation team (ET) analyzed a sample of 72 DoS evaluation reports from 2012 through 2016, using an evaluation quality checklist with 29 binary principal quality indicators (PQIs) and 15 supplemental quality indicators (SQIs), falling under the nine quality categories of: 1) executive summary; 2) report structure; 3) objectives and audiences; 4) EQs; 5) methodology; 6) findings; 7) conclusions; 8) recommendations; and 9) annexes. For the performance evaluation phase, the ET administered an online survey to DoS evaluation staff, which was answered by 66 persons from 25 DoS bureaus and independent offices. The survey was followed by 28 key informant interviews (KIIs) and two focus group discussions (FGDs), with a sample of 38 purposively selected staff and management personnel from 25 bureaus and independent offices that use FA funds. The ET analyzed and triangulated the data collected from each of these methods to answer the EQs and make recommendations to F for supporting DoS' evaluation efforts going forward.

Limitations to the evaluation methods include sampling bias, owing to a smaller than expected sample of DoS evaluation reports and survey response rate; key informants with limited knowledge and experience;

recall bias; halo bias; and reliance on subjective perceptions of DoS staff. The evaluation design compensated for these limitations through: 1) the triangulation of evidence; 2) actions to increase the quantitative and qualitative samples; and 3) the use of best practices in data collection, data analysis, and reporting.

Findings and Conclusions

EQ 1: What does the structure and content of evaluation reports tell us about the strengths and weaknesses of methodology, design, and analysis in evaluations?

Findings: A nearly equal number of reviewed evaluation reports were rated as fair (34) or good (30), while relatively few reports (8) were rated as poor. While the meta-evaluation sample did not allow a year-by-year assessment of report quality, the DoS staff interviewed believed the quality of evaluation reports has improved since the issuance of the Evaluation Policy in 2012, particularly in terms of creating better evaluation Statements of Work (SOWs), doing a better job identifying EQs, and providing more useful recommendations. Finally, the quality of single-country and functional bureau evaluation reports was higher than multi-country and regional bureau evaluation reports. In terms of specific quality categories and indicators, the meta-evaluation generated the following findings.

Executive Summary: The evaluation reports did a good job including executive summaries, but the executive summaries often lacked information about basic evaluation components, particularly evaluation conclusions and, to a lesser extent, findings and recommendations.

Report Structure: The evaluation reports did a good job organizing and presenting information, including organizing the report around EQs and covering key report components. The reports were consistently readable and accessible to lay audiences and slightly longer on average than DoS guidelines. At the same time, a sufficient number of reports either omitted conclusions or failed to differentiate conclusions from findings.

Evaluation Objectives and Audiences: The evaluation reports did a good job describing the evaluation's objectives and a poor job describing the evaluation's audience.

Evaluation Questions: The evaluation reports did a good job stating the key EQs and keeping the number of EQs within DoS' guidelines. The few reports that also included sub-EQs kept their number relatively low. Reports that included the evaluation SOW in the annexes did a fair job matching the EQs in the report to those in the SOW. Reports where EQs did not match did a poor job explaining why.

Methodology: The evaluation reports did a good job describing the evaluation methodology, methodological limitations and quantitative and qualitative sample sizes, but they only did a fair job describing the sampling and data analysis methods for both quantitative and qualitative data.

Findings: The evaluation reports did a good job drawing on all data collection methods in presenting the findings but largely failed to address alternative explanations or differential effects on women and different social groups.

Conclusions: The evaluation reports did a good job presenting conclusions that followed from the findings in those reports where conclusions were clearly differentiated from the findings.

Recommendations: The evaluation reports did a good job presenting specific recommendations that followed from the findings and conclusions but did a poor job presenting actionable recommendations.

Annexes: The reports did a good job including important information in the report annexes (SOW, bibliography, data collection instruments, list of persons interviewed) with the exception of evaluation teams' conflict of interest declarations.

Conclusions: The DoS evaluation reports reviewed for the meta-evaluation were, for the most part, of fair quality. While there were also a relatively large number of good reports, the number of fair or poor reports exceeded the number of good reports by a substantial margin, indicating significant room remains to improve the quality of DoS evaluation reports. Areas in which evaluation reports were relatively strong included discussing the evaluation methodology, drawing on all data collection methods, readability and accessibility to lay audiences, and describing methodological limitations. Areas in which evaluation reports were relatively weak included distinguishing findings from conclusions, providing actionable recommendations, describing both quantitative and qualitative sampling and data analysis methods, matching EQs in the report to those in the evaluation SOW, discussing gender and social effects, and considering alternative explanations. Subjectively, the overall quality of evaluation reports in the Department has improved since the Evaluation Policy was issued in 2012, owing to a combination of own initiative taken by bureaus and offices, motivation provided by the Evaluation Policy, and F technical and other assistance.

Different report characteristics did not, for the most part, explain the reports' relative quality, either overall, by quality category, or by individual PQIs. Notwithstanding, the relative disadvantages possessed by regional bureaus in terms of staffing, funding, capacity, and types of programs supported appear to have contributed to a lower overall quality of evaluation reports compared to reports produced by functional bureaus.

EQ 2: To what degree are FA evaluation findings and recommendations being used by bureaus, and what factors contribute to their use or non-use?

Findings: While stakeholders strongly support the importance of doing and using evaluations in theory, actual support, as evidenced by supportive actions, varies widely. DoS bureaus and independent offices only use evaluation findings and recommendations to "some extent" and for a variety of purposes, including to improve program implementation, contribute to bureau learning, inform program design and strategy, support change management, and question basic assumptions. The use of evaluations with bureaus and offices is, on the whole, more sporadic than systemic.

The primary factors hindering evaluation use include a lack of leadership support, unactionable recommendations, limited funding, low interest in evaluations, inadequate dissemination of evaluation results, failure to tie evaluations to bureau strategy, fear that evaluations will be used for punitive purposes, absence of a formal knowledge management system with the Department, and poor quality of the evaluation designs as reflected, for example, in their SOWs, data collection methodologies, work plans, understanding of project context, or technical grasp of the relevant subject matter.

Evaluation resources and technical assistance (TA) provided by F to bureaus and offices have provided critical support in helping certain bureaus and offices move up the evaluation learning curve.

Conclusions: The Evaluation Policy appears to have been effective at the margin in: putting evaluations on the Department's agenda; motivating increased attention to and implementation of evaluations; and establishing a framework for institutionalizing evaluation practice within the Department compared to the situation beforehand. The theoretical importance attached to evaluations within bureaus and independent offices is not matched by their practical importance, as measured by actual actions. At the same time, evaluation use in the Department is inconsistent and has yet to be institutionalized across bureaus and independent offices, albeit with some exceptions. Where evaluations are used, they are used for a variety of purposes, with the most common purposes being to inform program design and improve program implementation. There is no consensus as to which of the factors identified in the findings have played the biggest roles in hindering evaluation use.

Regarding knowledge management, F's efforts to increase the dissemination of evaluation results via the evaluation registry and evaluation CoP fall short of creating the systematic knowledge management processes needed. Further, the fear that poor evaluation results will be used for punitive purposes, particularly to reduce program funding, is an important and understandable concern. Any campaign to change the Department's evaluation culture must confront this issue better than what has been done to date. While F has developed a suite of tools addressing the range of factors hindering evaluation use, it has not instituted a process to instruct BECs and other DoS staff on how to use them or provided many practical examples of their use. Thus, BECs and other DoS staff who need these tools the most, and who are the most likely to use them, can find it difficult to use or adapt them.

EQ 3: To what extent have bureaus met the baseline requirements of the policy for planned and completed evaluations and staffing and utilizing the BEC position?

Findings: Each year, from 2014–2016, about 90 percent of bureaus and independent offices complied with the DoS Evaluation Policy by completing at least one evaluation per year or having at least one evaluation ongoing at the end of the year. In terms of provisions beyond the one-per-year requirement, less than half of key DoS staff interviewed believe their respective bureaus and independent offices are compliant with the Evaluation Policy. Overall, compliance with the Evaluation Policy varies across bureaus and independent offices.

While all bureaus and independent offices have staffed and are using the BEC position, the roles and responsibilities of BECs, the amount of time BECs spend on evaluation activities, and the experience and capacity of BECs vary significantly across DoS bureaus and independent offices. BECs range from highly experienced, visible, well resourced, supported, and proactive to inexperienced, inconspicuous, poorly resourced and supported, and reactive, without enough time or capacity to promote evaluations. The majority of the Department evaluation staff independently consult with their BECs on evaluation issues, with most of them finding the consultations useful. However, several do not consult with their BECs or do not find the consultations useful.

The above discrepancies, together with a lack of senior leadership buy-in, constitute primary challenges to achieving department-wide compliance with the Evaluation Policy. Other challenges include a lack of monitoring and evaluation (M&E) staff and technical expertise among existing staff, funding constraints, programs that do not easily lend themselves to evaluation, and bureaucratic external contracting obstacles. Overall, and owing in large part to the same challenges, evaluations have yet to play an important role in informing bureau policy, programming, and decision-making, and most BECs do not regularly engage bureau leadership, with some exceptions on a case-by-case basis.

Conclusions: Overall, familiarity with and support of the DoS Evaluation Policy within the Department is relatively high but far from universal in both cases. Thus, notwithstanding the efforts F has made to this point, there remains further work to be done to ensure that information about the Evaluation Policy penetrates to and is internalized by all staff and leadership levels within the Department. Bureaus and independent offices are complying with the requirement to conduct at least one evaluation per year. Perceptions of DoS staff about bureau and office compliance with other requirements of the Policy, on the other hand, vary widely in part due to lingering confusion about how the Evaluation Policy requirements apply to them and what it requires.

The wide variety in terms of BEC experience and capacity, responsibilities, time allocation, resources, and leadership and staff support, create serious challenges for achieving department-wide compliance with the requirements of the DoS Evaluation Policy and for institutionalizing evaluation practice within the Department. Other key challenges in meeting the Evaluation Policy's requirements are a lack of technical expertise, funding constraints, bureau programs that do not lend themselves to evaluation, bureaucratic

obstacles, and a lack of senior bureau leadership buy-in. The lack of senior leadership buy-in is a critical challenge that requires a planned and formal response beyond what has been done to date.

The Evaluation Policy and its implementation have had a significant impact on how bureaus and offices approach evaluation. Evidence for this includes improved staff M&E capacity, an increased number of evaluations being done, an increased focus on using recommendations for program learning and improvements, the institutionalization of the BEC position, the wider availability of evaluation tools and resources, and (despite challenges noted above) increased buy-in from senior bureau leadership compared to before the Policy.

EQ 4: To what degree have bureaus used F’s capacity building tools, classes, papers, and other technical assistance to increase evaluation capacity and improve evaluation?

Findings: Most DoS evaluation staff have used one or more of the F evaluation tools. F tools rated by more than half of evaluation staff to be either very or somewhat useful include the independent government cost estimate (IGCE), on-line resources, SOW toolkit, and evaluation design matrix. F tools rated by less than half of evaluation staff to be either very or somewhat useful include the evaluation wiki, newsletter, low/no cost evaluations, suggestions for reviewing reports, and dissemination summary template. The SOW toolkit was rated as the most useful tool, while low/no cost evaluations were rated as the least useful tool. The demand for the F-designed/BP-funded evaluation courses is high, and evaluation staff generally believe the courses have been effective in building DoS’ evaluation capacity.

Most DoS evaluation staff have consulted directly with F and participated in the evaluation CoP. The large majority of DoS evaluation staff that have consulted with or received TA from F have found the consultations and TA to be useful. On the other hand, participation in the CoP has not been consistent across evaluation staff and is largely limited to participating in CoP meetings. Notwithstanding evaluation staff generally acknowledge certain benefits of CoP participation, the perceived value-added of the CoP to their jobs varies considerably. While Department evaluation staff generally agree that it would be valuable to augment the existing suite of F evaluation tools, there is no general agreement as to what those new tools should be, aside from providing tools designed for the specific needs of users.

Conclusions: The large majority of Department evaluation staff are familiar with the F evaluation tools and have used one or more of the tools; however, a sizeable number of evaluation staff find the tools to be of limited usefulness. This lack of consensus about the usefulness of F tools reflects the wide variety of experience and capacity found among Department evaluation staff and the difficulty of designing evaluation tools that appeal to an audience with diverse needs. Department evaluation staff, moreover, are largely left to their own initiative to seek out and use or adapt the F tools. The lack of direct training in the tools as well as the lack of practical examples in how to use/adapt them hinder their use, in the latter case, the result of a Department culture that discourages sharing and cooperating across bureaus and offices.

DoS evaluation staff are generally consulting directly with F and attending CoP meetings. The value-added of CoP participation, however, is limited by two factors. The first is the difficulty involved in structuring CoP meetings to meet the diverse needs of Department evaluation staff. The second is the absence of formal (or structured) networking opportunities it offers to participants. F’s assistance via direct consultants and TA has been effective and offers an important and useful supplement to the standardized nature of the F evaluation tools and F/BP evaluation courses. For the relatively few evaluation staff that did not find F consultations or TA useful, no reasons emerged with sufficient consistency to suggest any underlying problems.

The demand for the F and BP evaluation courses is high, reflecting their perceived value among Department evaluation staff. At the same time, their perceived value is limited by an overly standardized

approach - ill-suited at times to participants' diverse needs - that fails to engender creative application of evaluation practices. The failure to incorporate F's evaluation tools into the courses is an important oversight that limits their adoption by Department evaluation staff. There exists demand for F to augment its suite of evaluation tools, but this demand exists largely in a general form. Aside from a desire for F to design tools that are better targeted to users' specific needs, there is no consensus among stakeholders as to what specific forms these augmented tools should take.

EQ 5: Looking forward to implementation of FATAA, what areas are likely to require more attention or focus from F?

Findings: Additional F support is required to help bureaus implement FATAA and, by extension, the revised Evaluation Policy, particularly through promoting buy-in and understanding of bureau leadership, linking evaluation results to decision-making, facilitating the creation of a department-wide M&E support team, providing more and better guidance on how to comply with FATAA, providing guidance to bureaus on how to address common objections or concerns about evaluations, and promoting greater cross-bureau collaboration, including mentoring by more experienced and more capable evaluation staff to less experienced and less capable evaluation staff.

Conclusions: There is a need for more support from F in multiple areas to help bureaus and independent offices achieve full compliance with FATAA. Priority areas in which additional F support is required are described in the findings.

Recommendations

EQ 1: To improve the methodology, design, and analysis in evaluation reports, F should:

1. Provide more detailed evaluation report guidance. F should provide more detailed guidance to BECs and other DoS evaluation staff on quality criteria for evaluation reports. This guidance should be structured to serve as a guide to both preparing and reviewing evaluation reports. While F has provided guidance on the evaluation content and quality, this guidance is at a fairly general level and could benefit from increased scope and specificity.
2. Link to examples of good quality reports. F should embed links or references to examples of good quality reports - or reports that demonstrate good quality with regards to certain quality criteria - drawn from its existing database of reports in all of its guidance documents, so relevant DoS staff can see examples of good practice.
3. Establish a process for the review of draft evaluation reports. F should consider establishing a process for reviewing draft evaluation reports that taps into existing resources and evaluation expertise within the Department.
4. Work with partners to improve evaluation reports and evaluation capacity. F should work with external contractors, including sharing existing report preparation guidance, to increase their capacity for creating quality evaluation reports with specific and actionable recommendations.

EQ 2: To improve the use of evaluation findings and recommendations by bureaus, F should:

1. Develop a strategy for engaging bureau leadership. F should develop a strategy to increase leadership buy-in for evaluations that: 1) identifies existing opportunities to engage directly with leadership; 2) creates a network of change agents; and 3) advocates for the incorporation of evaluation concepts into Foreign Service Institute (FSI) training for senior leadership.
2. Develop a knowledge management strategy. F should develop a strategy to implement a department-wide knowledge management strategy, including the dissemination, management, and use of evaluation results and lessons learned within bureaus.
3. Provide guidance on creating learning products to enhance outreach of key evaluation takeaways. F should provide additional guidance to BECs and external contractors on strategies for disseminating evaluation results that increase accessibility and readership among evaluation stakeholders, such as

short summaries of evaluation findings or recommendations, summary charts or graphics, highlights of successful interventions or use of evaluation recommendations, videos, etc.

EQ 3: To support bureaus meeting the requirements for evaluation and the BEC position, F should:

1. Supplement existing guidance on compliance with the DoS Evaluation Policy. F should supplement existing policy guidance with easily digestible summary guidance in the form of a 1- or 2-page document that summarizes the policy and compliance requirements.
2. Provide additional guidance on BEC qualifications. F should provide additional guidance on desired BEC qualifications and their roles and responsibilities. This would augment existing guidance with more expansive and detailed language that helps bureaus and independent offices think more critically about the BEC's role.

EQ 4: To increase bureau evaluation capacity and improve evaluation, F should:

1. Integrate F evaluation tools into evaluation courses. F should integrate its evaluation tools into future evaluation training courses to introduce Department evaluation staff to the tools and instruct them in their use. F should further include a provision in the external contractor SOW that requires them to use the F tools in developing and delivering the training courses.
2. Seek opportunities to link evaluation staff to F evaluation resources. F should seek out and take advantage of opportunities to link DoS' evaluation staff to both internal and external resources with examples of good practice and practical tool application.
3. Enhance the benefits of CoP membership. F should seek to enhance the benefits of CoP membership through additional activities of interest to groups of members that are segmented by evaluation experience and capacity levels.
4. Bring BECs and other evaluation staff together for networking and joint learning. F should facilitate formal opportunities for BECs and other DoS evaluation staff to meet for the specific purposes of networking and joint learning. Examples include workshops, brown bags, happy hours, etc.
5. Develop BEC induction resources. F should develop a package of resources that it provides to new or inexperienced BECs for the purpose of orienting them and making sure they understand their jobs and where they can find resources or help.

EQ5: To support the implementation of FATAA, F should:

1. Offer evaluation courses targeted to different audiences. F should provide options for evaluation courses that are targeted to different audiences with different experience and capacity levels.
2. Facilitate peer mentoring among Department evaluation staff. F should help facilitate peer mentoring opportunities among Department evaluation staff, thereby offloading some of its TA to others within the Department who possess the capacity and willingness to serve as mentors. F can create the framework for peer mentoring and then act to help link supply to demand.

1.0 EVALUATION PURPOSE AND EVALUATION QUESTIONS

1.1 Evaluation Purpose and Audiences

The United States (U.S.) Department of State's (DoS) Office of U.S. Foreign Assistance Resources (F) commissioned an independent external evaluation of the implementation of the Department's Evaluation Policy. The goal of the evaluation was to take stock of the Department's experience with the evaluation of Foreign Assistance (FA)-funded programming from 2012 to the present for the purpose of informing F's decisions about how to further institutionalize evaluation of FA across the Department.

The evaluation was implemented by ME&A, located in Bethesda, MD. The purpose of the evaluation was to: 1) evaluate the quality of completed evaluations through an examination of their strengths and weaknesses on the basis of report content; 2) examine the effectiveness of completed evaluations by determining the extent to which evaluation findings and recommendations are used and what obstacles, if any, hinder their utilization; and 3) determine the extent to which the significant efforts undertaken by the DoS over the last several years to institutionalize evaluation have been effective in building up evaluation expertise and capacity in bureaus to enable them to implement the new Department policy and respond to the new requirements of Foreign Aid Transparency and Accountability Act (FATAA). The evaluation focused on evaluation practice in the 26 DoS bureaus and independent offices that have received FA funds.¹

The primary audience for the evaluation is F. Secondary audiences include the Bureau of Budget and Planning (BP), which oversees the Evaluation Policy for Diplomatic Engagement (DE)-funded programming and works closely with F; DoS' functional and regional bureaus that use FA funds; Congress; and the broader foreign assistance community.

1.2 Evaluation Questions

The evaluation answers the five evaluation questions (EQs) posed in the evaluation Statement of Work (SOW) (see Annex I), which fall under three evaluation themes: 1) strengths and weaknesses of evaluation reports; 2) use of evaluations; and 3) capacity development in bureaus. The five EQs are listed below.

Strengths and Weaknesses of Evaluation Reports

EQ1. What does the structure and content of evaluation reports tell us about the strengths and weaknesses of methodology, design, and analysis in evaluations?

Utilization of Evaluations

EQ2. To what degree are FA evaluation findings and recommendations being used by bureaus and what factors contribute to their use or non-use?

Capacity Development in Bureaus

EQ3. To what extent have bureaus met the baseline requirements of the policy for planned and completed evaluations and staffing, and utilizing the bureau evaluation coordinator (BEC) position?

¹ The number of bureaus and independent offices that receive FA funds varies from year to year but typically falls in the range of 22-26.

EQ4. To what degree have bureaus used F’s capacity building tools, classes, papers, and other technical assistance to increase evaluation capacity and improve evaluation?

EQ5. Looking forward to implementation of FATAA, what areas are likely to require more attention or focus from F?

2.0 BACKGROUND

F established the first department-wide Evaluation Policy in 2012, along with BP. The policy promoted evaluations to increase learning from programming, increase the effectiveness of programs, and provide data to support decision-making. F issued evaluation guidance as part of this process, provided support for evaluation in the Department (along with BP, which oversees implementation for DE-funded programming), and monitored implementation of the policy over two years. F and BP then developed a revised Evaluation Policy in 2015, and updated guidance and support for evaluation at that time. F developed a series of papers and tools to help bureaus and offices with the design, implementation, and utilization of evaluations, and updated older tools and templates. Following passage of FATAA, F, in conjunction with BP, revised the Evaluation Policy a third time to respond to the requirements of the legislation. Among other things, the revised policy added language that made bureaus and independent offices responsible for performance management by placing evaluation as part of a continuum of planning, managing, monitoring, evaluation, and learning. The updated policy was issued in November 2017. Because it addresses more than evaluation, F sought to rebrand the expanded version of the Evaluation Policy as “18 FAM 300,” which is the location where the policy language is codified in the DoS Foreign Affairs Manual (FAM).

Since 2013, F and BP have offered courses on managing evaluations, evaluation design, and data collection methods for the Department. Both F and BP have provided technical assistance to bureaus through consultations on planning for evaluations, developing SOWs, providing tools and reference materials, and reviewing draft reports, as well as organizing conferences, courses, and workshops. Per the policy, bureaus and independent offices have assigned BECs who support the planning and management of evaluations in their bureaus and offices. F and BP further manage an evaluation community of practice (CoP) for the Department, which meets monthly to discuss issues and share information. F also manages online resources to support bureaus with additional information on evaluation.

As a result of these developments and the requirements of the Evaluation Policy, bureaus have been increasingly undertaking evaluations of their programs. The functional and regional bureaus as well as independent offices that use FA funds have conducted more than 180 evaluations from 2012 through 2017.

3.0 EVALUATION DESIGN, METHODS, AND LIMITATIONS

3.1 Evaluation Design and Methods

The evaluation used a mixed-methods design consisting of a combination of quantitative and qualitative data collection methods. The design was developed to answer all the EQs via a process including preparation, data collection, analysis, and reporting. The information below is a relatively brief summary of the data collection and analysis methods used for the evaluation. (For a more detailed description of the evaluation methodology, see Annex VIII. For a tabular summary of the EQs and their corresponding data collection methods, see the evaluation matrix in Annex III.)

3.1.1 Inception Phase and Document Review

The first phase, or inception phase, began with a kick-off meeting with F to align expectations and understandings related to the evaluation. The evaluation team (ET) then reviewed documents provided by F and additional materials on meta-evaluation to develop the ET's understanding of the experience to date at DoS with regards to evaluation and to inform the evaluation design, including data collection and analysis methods. (See Annex II for a bibliography of documents reviewed for the evaluation.) During the inception phase, the ET developed and the F approved the evaluation design—including the data collection instruments (meta-evaluation checklist, online survey questionnaire, interview guides, and sampling plan)—and evaluation work plan.

3.1.2 Meta-Evaluation Phase

DoS Evaluation Policy Dissemination Requirements

In the second, or meta-evaluation, phase, the ET analyzed a sample of DoS' evaluation reports provided by F, covering the period 2012-2016. The DoS Evaluation Policy requires that F be given copies, through an evaluation registry, of all completed, unclassified evaluation reports for internal learning purposes and for publication on state.gov. If an evaluation is sensitive but unclassified, bureaus may submit a summary of the findings for publication. Classified evaluations are exempt from the policy.

Number of Evaluation Reports Made Available to the Evaluation Team

The SOW for this evaluation stipulated that the meta-evaluation would consist of a review of up to 124 evaluation reports; however, the ET's preliminary examination of the evaluation reports provided by F yielded a set of only 72 complete, comparable reports for the meta-evaluation. The rest of the reports were excluded from the sample because they were not evaluations or were duplicates, summaries of evaluations, reports that were subsumed in other evaluation reports, or evaluations that were conducted by organizations other than DoS.

The final set of 72 evaluation reports was significantly smaller than the 124 reports anticipated in the SOW and constituted 40 percent of the 181 evaluation reports completed by DoS bureaus and independent offices from 2012-2016. This figure includes 34 of 118 evaluations (29.7 percent) completed during 2012-2014, and 48 of 63 completed reports (76.2 percent) submitted during 2015-2016.²

The distribution of full evaluation reports reviewed for the meta-evaluation organized by their year of completion is shown in Table 1.

Table 1. Distribution of Evaluation Reports Reviewed for the Meta-Evaluation by Year of Completion

	2012	2013	2014	2015	2016	Total
Number reports reviewed	5	15	14	26	12	72
Number evaluations completed	79		39	32	31	181
Percentage reviewed/completed	38.0%		35.9%	81.3%	41.9%	39.8%

As can be seen there, the percentage of reports reviewed relative to reports completed ranged from a low of 35.9 percent in 2014 to a high of 81.3 percent in 2015. (The number of reports completed in each of 2012 and 2013 was not available; only the combined number of reports completed during 2012-2013 was available.) The number of evaluation reports reviewed for any given year does not necessarily reflect

² The figure of 181 DoS evaluation reports is based on values reported in the 2015 OIG report, "Review of Department of State Compliance with Program Evaluation Requirements" for February 2012-September 2014 and on values for 2015 and 2016 provided by F to the ET. The ET does not have information as to how many of these reports were unclassified, sensitive but unclassified, or classified.

anything particular for that year other than these were the evaluation reports submitted to F by the DoS bureaus and independent offices during that year.

Possible Explanations Why the Number of Evaluation Reports Was Lower than Anticipated

A few reasons help explain why the number of reports available for the meta-evaluation was lower than anticipated. First, F shared with the ET all the unclassified evaluation reports in its possession and was not aware of the issues with the reports discovered by the ET and described above. Second, except for a small number of bureaus that published evaluations on their state.gov pages, few bureaus shared evaluations internally or externally prior to the Evaluation Policy. Third, a relatively large number of evaluation reports are either classified or sensitive but unclassified. Fourth, not all bureaus are complying fully with the dissemination requirements in the Evaluation Policy. The ET was unable to obtain information to determine to what extent the third and fourth explanations are accurate.

The Meta-Evaluation Checklist

To conduct the meta-evaluation, the ET used a checklist scoring system consisting of 29 primary quality indicators (PQIs) and 15 secondary quality indicators (SQIs), grouped in nine broad scoring categories: 1) executive summary; 2) report structure; 3) evaluation objectives and audiences; 4) EQs; 5) methodology; 6) findings; 7) conclusions; 8) recommendations; and 9) annexes. The checklist used a dichotomous scoring system (Yes/No) to measure whether the evaluation report did or did not meet the relevant quality criteria in which a score of 1 indicated that the report met the relevant quality criteria, and a score of 0 indicated that the report did not meet the relevant quality criteria. The scores on each PQI were then summed to arrive at the total report quality score, with a maximum quality score of 29.

The 29 PQIs included in the meta-evaluation checklist were indicators that the ET determined were essential for a high-quality evaluation report. The 15 SQIs were added to the checklist to provide additional granularity for analytical purposes but were not used for calculating the report quality score. The 29 PQIs and 15 SQIs, and the scoring categories under which they fell, are shown in Table 2. (See Annex IV for a copy of the checklist and Annex IX for a summary of meta-evaluation results.)

Table 2. Number of Principal and Supplemental Checklist Indicators by Scoring Category

Scoring Category	Number of PQIs	Number of SQIs
Executive Summary	2	2
Report Structure	4	2
Evaluation Objectives & Audience	2	0
EQs	3	3
Methodology	5	6
Findings	3	1
Conclusions	2	0
Recommendations	3	1
Annexes	5	0
Total	29	15

Developing and Refining the Meta-Evaluation Checklist

In selecting indicators for the meta-evaluation checklist, the ET sought to focus on report quality criteria included in the January 2016 DoS evaluation guidance document, “Guidance for Planning and Conducting Evaluations at the Department of State” (2016 DoS Evaluation Guidance). Of the 29 PQIs included in the checklist, the ET extracted 24 directly from the 2016 DoS Evaluation Guidance. (The meta-evaluation checklist found in Annex IV indicates the page in the 2016 DoS Evaluation Guidance from which the ET extracted the relevant PQI along with a rationale for including the indicator and supporting language taken from the guidance document.) The ET derived the remaining five PQIs from other sources, including meta-

evaluations conducted by MSI for USAID and by the Government Accountability Office (GAO) for the President's Emergency Plan for AIDS Relief (PEPFAR), and from their own experience.³

After developing the checklist, the ET conducted two pilot tests to identify problems with the checklist and ensure that all coding team members used the checklist in identical ways. After conducting each pilot test, the ET met to compare their assessments, which they used to answer questions, resolve problems, clarify and strengthen scoring guidelines, and make further refinements to the checklist indicators and scoring sheet. At the end of this process, the ET finalized the checklist indicators, scoring system, and scoring guidelines, and launched the meta-evaluation. Three members of the ET conducted the meta-evaluation over a period of two weeks, dividing up the 72 evaluation reports more or less equally among them.

After the coding team had completed scoring all 72 evaluation reports, the ET noted that a number of the PQIs were judged to have met the quality criteria in 100 percent or near 100 percent of all reports using the dichotomous scoring system. On investigating this result further, the ET noted that four of these PQIs—"report is readable," "recommendations are supported by the findings and conclusions," "recommendations are specific," and "recommendations are actionable"—involved a more than typical degree of subjective judgment, suggesting that it would be appropriate to re-score them using a scalar scoring system that allowed the ET to uncover greater variations in the reports, which were not evident using the dichotomous scoring system.⁴ Thus, the ET undertook to re-score these four PQIs using the following 5-point rating scale: 1=Poor, 2=Weak, 3=Fair, 4=Good, and 5=Excellent. This rating scale was used for all meta-evaluation analysis and the resulting findings are presented below. (See Annex VIII for a more detailed description of the ET's rationale and method for re-scoring the relevant four PQIs.)

3.1.3 Performance Evaluation Phase

In the third, or performance evaluation, phase, the ET undertook two major research data collection activities to answer EQs 2-5 on the utilization of evaluation findings and recommendations and building evaluation capacity in bureaus. Activities included: 1) an online survey using the SurveyMonkey platform; and 2) key informant interviews (KIIs) and focus group discussions (FGDs) with purposively selected staff and managers from a cross-section of DoS bureaus and independent offices.

Online Survey

Based on the meta-evaluation results and after discussion with F, the ET revised the online survey questionnaire developed during the inception phase and received final approval to commence with the survey. The sampling frame for the online survey was the email list of 213 FA staff participating in the DoS' evaluation CoP. The survey was originally scheduled to occur during the two weeks from December 11-22, 2017. Prior to the beginning of the survey, F composed and sent an email letter of introduction explaining the purpose for the survey and providing instructions on how to access it at the SurveyMonkey site. A week after sending out the initial letter of introduction, F sent out a follow-up reminder.

At the end of the scheduled two-week period for the online survey, the ET and F decided to extend the time period an additional two weeks, and F sent out a third follow-up email announcing that the timeframe for completing the survey had been extended and again asking recipients to complete the survey. In the end, 66 persons from each of the 25 DoS bureaus and independent offices receiving FA

³ Molly Hageboeck, Micah Frumkin, and Stephanie Monschein, (2013), "Meta-Evaluation of Quality and Coverage of USAID Evaluations 2009-2012," Management Systems International (MSI); Government Accountability Office, (2012), "Agencies Can Enhance Evaluation Quality, Planning, and Dissemination," President's Emergency Plan for Aids Relief (PEPFAR)."

⁴ In contrast, the ET concluded that the dichotomous scoring system was appropriate for the other PQIs that met the quality criteria in or near 100 percent of reports, which included "report describes methodology," "findings draw on all data collection methods," and "report describes objectives."

funds responded to the survey, for a 31 percent response rate.

The survey included 54 questions on respondent's demographic characteristics; the importance of doing evaluations; how bureaus have used evaluations; factors hindering evaluation use; familiarity, relevance, and adherence to the DoS Evaluation Policy; and familiarity, use, and usefulness of F evaluation tools, technical assistance (TA), and the CoP. (See Annex V for a copy of the online survey questionnaire, Annex X for a comprehensive summary of the online survey response, and Annex XI for the results of the analysis testing relationships between the online survey results and different respondent categories.)

KIIs and FGDs

Based on the meta-evaluation and online survey results, the ET revised the draft KII and FGD interview guides created during the inception phase and developed separate interview guides for three groups of stakeholders: 1) F and BP; 2) BECs; and 3) bureau leadership (e.g., Deputy Assistant Secretary, or DAS). F provided a list of evaluation key informants in each of the above three stakeholder groups whom the ET interviewed over the three-week period of January 15-February 2, 2018.

As it was not possible in the evaluation timeframe to interview all evaluation staff within the bureaus and independent offices receiving DoS FA funding, the ET worked with its counterparts at F to interview as comprehensive a sample of BECs in the Department as possible, in addition to a purposive sample of other bureau key informants. In the end, the ET conducted 28 KIIs and two FGDs with 30 of 36 BECs and back-up BECs, and six bureau leaders. (See Annex VI for a copy of the KII and FGD guides, and Annex VII for a comprehensive list of persons participating in KIIs and FGDs along with their DoS affiliation and position.)

3.1.4 Data Analysis Phase

In the fourth, or data analysis, phase, the ET synthesized and analyzed the data to identify key findings, conclusions, and recommendations. For the quantitative data collected via the meta-evaluation and online survey, the ET conducted all data analysis using either Excel or the SPSS data analysis software. Quantitative data analysis consisted primarily of generating frequencies and descriptives for the meta-evaluation and survey data, and conducting statistical tests, including chi-square tests of association, t-tests of the difference between two means, and analysis of variance (ANOVA) of the difference between three or more means. For the qualitative analysis, the ET conducted content analysis of the secondary documents and KII and FGD notes using the Atlas.ti data analysis software.

3.1.5 Report Writing Phase

In the fifth and final phase, the ET prepared the evaluation report. Individual team members assumed responsibility for undertaking different aspects of the data analysis and contributing narrative sections of the evaluation report. Draft sections were shared between ET members and discussed in conference calls and via email. The ET synthesized the findings, conclusions, and recommendations from each team member through this process to reach a set of agreed-on conclusions and recommendations from the accumulated findings.

3.2 Evaluation Team

The ET consisted of the following five individuals:

- Dr. Gary Woller, Team Leader. Dr. Woller was responsible for overseeing, coordinating, and participating in all activities related to the evaluation, and for ensuring the production and completion of quality deliverables in a professional manner, in compliance with the evaluation SOW.
- Dr. Lawrence Robertson, Senior Evaluation Specialist. Dr. Robertson participated in the development of the study design and work plan, supervised the meta-evaluation of the DoS evaluation reports, conducted KIIs, and contributed content to the final report.

- Christopher Coffman, Evaluation Specialist. Mr. Coffman participated in the development of the study design and work plan, coded the DoS evaluation reports for the meta-evaluation, conducted KIIs and FGDs, performed data analysis, and contributed content to the final report.
- Dereje Gebreegziabher, Research Analyst. Mr. Gebreegziabher participated in the development of the study design and work plan, coded the DoS evaluation reports for the meta-evaluation, conducted KIIs and FGDs, performed data analysis, and contributed content to the final report.
- Andres Rueda, Research Analyst. Mr. Rueda participated in the development of the study design and work plan, coded the DoS evaluation reports for the meta-evaluation, and contributed content to the final report.

3.3 Methodological Limitations

Potential methodological limitations to this evaluation, and the methods taken by the ET to mitigate them, are discussed below.

Sampling bias: Sampling bias refers to potential biases introduced by the selection of individuals for analysis in such a way that the sample obtained may not be representative of the underlying target population. For the meta-evaluation the primary potential source of sampling bias was the low number of reports provided to the ET by F relative to the 124 reports anticipated in the evaluation SOW, and the 181 evaluation reports completed by DoS bureaus and independent offices during 2012-2016. There is no way to ascertain whether and the extent to which the 72 reports reviewed for the meta-evaluation are representative of the evaluation reports produced by bureaus and independent offices over that period. If we were to assume that the 72 evaluation reports constituted a representative sample of reports completed during 2012-2014, similar to what would be achieved using simple random sampling, the sample would yield a confidence interval (margin of error) of ± 9 percentage points with a 95 percent confidence level.

In addition, the relatively low number of reports included for any of the five years covered by the meta-evaluation makes any attempt to assess trends in report quality using year-by-year comparisons a tenuous endeavor. For this reason, the ET, with F's consent, opted not to conduct a trend analysis of evaluation report quality for this report.

For the online survey, the primary risk of sampling bias was the non-random, low response rate achieved (66 of 213), despite multiple invitation emails sent out by F over a three-week period. If we were to assume that the 66 survey responses constituted a representative sample of the underlying population similar to what would be achieved using simple random sampling, the sample size would yield a confidence interval of ± 11 percentage points with a 95 percent confidence level.

Finally, for the KIIs and FGDs, the primary risk of sampling bias was the non-random selection process of key informants. This was not an issue for the 36 BECs and back-up BECs in the Department (of whom the ET interviewed 30), but it was an issue for bureau leadership, of whom the ET only interviewed six.

Information sources with limited experience in evaluation: DoS personnel are frequently rotated from position to position, meaning that personnel contacted to provide information for the evaluation may or may not possess the background or experience to provide informed answers to the survey questions or responses to the KIIs and FGDs.

Recall bias: Respondents may remember events in the past less clearly than more recent events. Since a number of interview questions focused on issues that took place in the past, responses may have been less accurate about events from the more distant past relative to more recent times.

Halo bias: Social science research has demonstrated that the tendency of respondents to underreport undesirable or uncomfortable views and behaviors is common, and that respondents may alter responses to questions to report more socially accepted views and behaviors. The extent to which respondents are

prepared to reveal their true opinions may also vary for questions that call on the respondents to assess their own performance, the performance of their colleagues, or their assessments of their management and supervisors.

Reliance on subjective perceptions: The answers to EQs 2-5 rely heavily on the subjective perceptions and recollections of KII, FGD, and online survey respondents, as opposed to objectively verifiable evidence.

Inter-rater reliability: For reasons related to time constraints, the ET used a three-person coding team to review evaluation reports and assign them scores for the meta-evaluation. The risk of this approach is that different coders may apply different criteria or employ different subjective judgments in scoring the reports, resulting in low inter-rater reliability and thus calling into question the validity of the meta-evaluation findings.

3.3.1 Actions Taken to Mitigate Methodological Limitations

The ET developed the evaluation design to compensate for the above limitations through the following methods:

Triangulation of evidence: The ET triangulated the evidence collected through the different methods and from different qualitative and quantitative data sources, which reduced the risk that biases inherent in any one data source unduly influence the evaluation's findings and conclusions.

Actions to increase/maximize response rates to online survey and KII and FGD requests: To stimulate participation and reduce the risk of sampling biases from non-participation, the ET received substantial support from its F counterparts to solicit DoS staff participation in the online survey, KIIs, and FGDs. For the online survey, F sent out the invitation to the entire FA CoP membership list that served as the sampling frame for the survey, as well as three follow-up emails to the same recipients. This enabled staff to be invited through an official state.gov email address, which the ET and F believed would increase the credibility of the request.

F staff also used DoS' systems to request and set up KIIs with BECs, deputy assistant secretaries (DASs), and key DoS evaluation staff. Again, invitations from a state.gov email address were expected to be more compelling than if requests for a meeting came from ME&A, an organization external to the Department. Of all the meetings requested for the KIIs, fewer than five invitees in the Department did not agree to a meeting with the ET over the two-and-a-half-week period for interviews and focus groups in the Department.

Use of best practices in data collection: The ET applied best practices in quantitative and qualitative data collection and analysis to maximize the valid use of these data and minimize biases. The ET established and tested data collection protocols for each of the data collection instruments used in the evaluation and met weekly (either in-person or virtually) to review the activities, compare notes and experiences, note and address problems, and make adjustments as appropriate to the following week's activities. The ET also held a weekly telephone conference with F to ensure close collaboration throughout the evaluation.

Use of best practices in data analysis and reporting: The ET conducted systemic data analysis using well-established analysis methods supported by advanced qualitative and quantitative data analysis software. The use of SPSS for quantitative analysis and Atlas.ti for qualitative analysis added rigor to the data analysis process, reduced the risk of biases from the work of any individual team members, and increased the credibility of the findings. In conducting the data analysis and reporting the results in this report, moreover, the ET has focused, for the large part, on findings based on patterns in the data that recurred with relative frequency.

Pilot testing and ongoing team consultations: To reduce the risk of low inter-rater reliability, the entire ET participated in a pilot test of the meta-evaluation checklist and adjusted the checklist and scoring criteria

using the process described above. During the implementation of the meta-evaluation, the coding team remained in constant communication during which they identified and resolved issues as they arose, even to the point of going back and re-scoring reports and/or indicators should the situation merit it. After the completion of the meta-evaluation, the ET ran a number of statistical tests to determine whether significant differences in scoring patterns emerged among members of the coding team. These tests found no statistically significant evidence that scores differed among the coders.

4.0 FINDINGS AND CONCLUSIONS

This section presents the evaluation findings and conclusions related to each of the five primary EQs. The findings related to EQ 1 draw primarily on the results of the meta-evaluation, while the findings for EQs 2-5 draw primarily on the online survey, KIIs, and FGDs, although answering each question will draw on a range of the data collection methods.

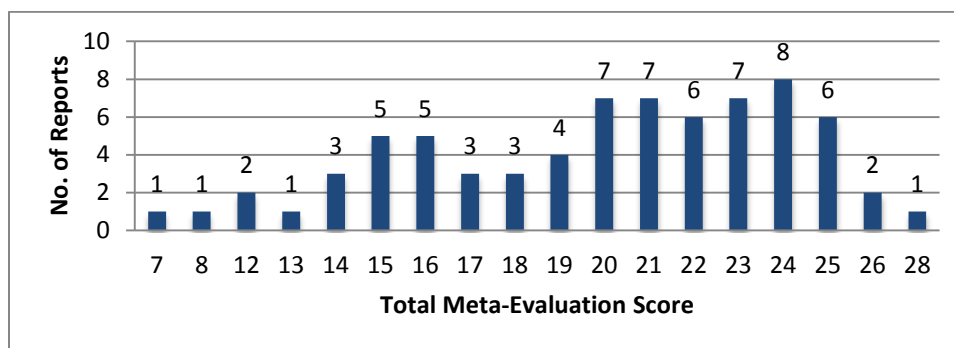
4.1 EQ 1: What does the structure and content of evaluation reports tell us about the strengths and weaknesses of methodology, design, and analysis in evaluations?

Overall Findings
<p>A nearly equal number of reviewed evaluation reports were rated as fair (34) or good (30), while relatively few reports (8) were rated as poor. DoS staff interviewed believed the quality of evaluation reports has improved since the issuance of the Evaluation Policy in 2012, particularly in terms of creating better evaluation SOWs, doing a better job identifying EQs, and providing more useful recommendations. Finally, the quality of single-country and functional bureau evaluation reports was higher than multi-country and regional bureau evaluation reports. Specific findings included the following:</p> <ul style="list-style-type: none"> • The reports did a good job including an executive summary. • The reports did a good job presenting information and were readable for lay audiences. A number of reports, however, either omitted conclusions or failed to differentiate conclusions from findings. • The reports did a good job describing the evaluation’s objectives and a poor job describing the evaluation’s audiences. • The reports did a good job stating the key EQs and keeping the number of EQs within DoS guidelines, and did a fair job matching the EQs in the report to those in the SOW. • The reports did a good job describing the evaluation methodology, methodological limitations, and sample size and did a fair job describing the sampling and data analysis methods. • The reports did a good job drawing on all data collection methods but did a poor job addressing alternative explanations or differential effects on women and different social groups. • The reports did a good job presenting conclusions that followed from the findings in those reports where conclusions were clearly differentiated from the findings. • The reports did a good job presenting specific recommendations that followed from the findings and conclusions but did a poor job presenting actionable recommendations. • The reports did good job including important information in the report annexes with the exception of evaluation teams’ conflict of interest declarations.

4.1.1 Overall Evaluation Report Quality

The presentation of findings for EQ 1 begins in Figure 1 with the presentation of the total meta-evaluation scores as measured by a report’s total score summed across all 29 PQIs.

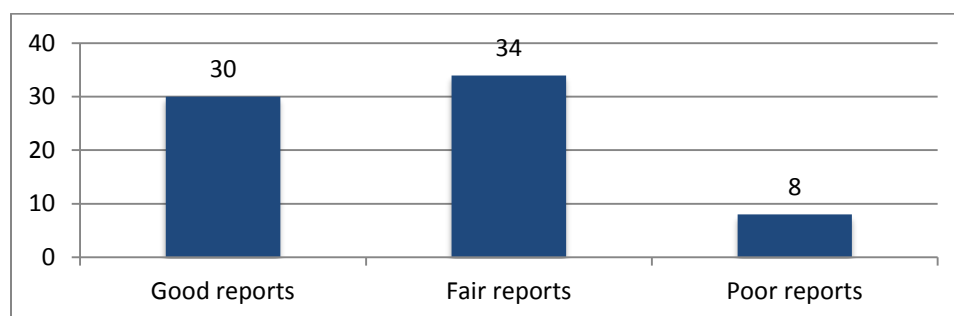
Figure 1. Distribution of Overall Meta-Evaluation Scores



As seen in Figure 1, total scores ranged from a low of 7 to a high of 28. The average total score for all 72 reports was a 19.9, with a median score of 21 and modal score of 24. (The X-axis in Figure 1 shows the range of overall scores earned, and the Y-axis shows the number of reports that earned a particular score. For example, three reports earned a score of 14, eight reports earned a score of 24, etc.)

What constitutes a “good” report based on its overall report score is a largely subjective determination; however, for this report, the ET adopted the standard that a “good” report earns at least 75 percent (22) of the total 29 points possible, while a “fair” report earns from 50-74 percent (15-21) of points possible, and a “poor” report earns less than 50 percent (15 <) of points possible. Using these criteria, 30 (41.7 percent) of the 72 reports were rated to be good, 34 (47.2 percent) were rated to be fair, and 8 (11.1 percent) of the reports were rated to be poor (see Figure 2).

Figure 2. Number of Good, Fair, and Poor Evaluation Reports



KIIs reported a wide range of opinions about the quality of evaluation reports in their bureaus; however, there was a general perception among key informants that the overall quality of reports has been improving since the issuance of the DoS Evaluation Policy in 2012, owing in part to the bureaus’ own initiative to improve its evaluation capacity in response to the Evaluation Policy and in part due to F’s assistance. (Unfortunately, this perception could not be confirmed or disconfirmed in the meta-analysis due to the structure of the report sample.) Referring to the general quality of evaluation reports, one BEC noted, *“Evaluation quality has improved over time. Now our external evaluators have a better understanding of our programming and they perform better, which has been facilitated by better program management performance and a more informed approach to examining the merits of proposals from external bidders.”*

When speaking about the quality of evaluation reports, key informants invoked a number of criteria. One common theme expressed in the KIIs was that because of an increase in technical capacity, bureaus are writing better SOWs with stronger EQs, properly reviewing evaluation plans and methodologies, and developing better results indicators. One KII, for example, reported that, *“For those that aren’t great, it is often because they are missing something that is usually a result of the program team not conveying*

the right information in the design and SOW, not the fault of the contractors.” According to another key informant, “Twenty to 30 percent [of reports] are not very good, often because they did not understand the questions well and did not generate the type of data that was needed. In many cases this was not their fault but was because the scope of work did not clearly identify the issues and the evaluation questions and what was wanted.” He continued: “The problem was more with the evaluation SOWs in that they were written in a way that allowed the evaluation team to do the bare minimum with standard methods and a lack of creativity. The methodology sections in the reports were not articulated in a way that provided confidence in how they were conducted or how participants were selected.”

Another prominent criterion mentioned by key informants was the usefulness of the evaluation reports, particularly the reports’ recommendations. One key informant, for example, described the success of a recent process evaluation, noting that the bureau has already incorporated half of the recommendations and is considering incorporating the other half, pending input from other stakeholders. He added, “We can point to tangible program improvements that have come as a result of the evaluation reports, and we are using evaluations to a high degree.” Yet another key informant estimated that 70 to 80 percent of the evaluation reports have been good in that they included useful information that answered the EQs using evidence and with actionable recommendations.

Conclusion 1
In terms of their overall quality, a moderate majority of DoS evaluation reports were of either fair or poor quality, suggesting that significant room remains to improve the overall quality of DoS evaluation reports. Subjectively, the overall quality of evaluation reports in the Department has improved since the Evaluation Policy was issued in 2012, owing to a combination of own initiative taken by bureaus and offices, motivation provided by the Evaluation Policy, and F’s technical and other assistance.

4.1.2 Evaluation Report Quality by Quality Category

At a higher level of disaggregation, the ET used three quality criteria—good, fair, and poor—to assess report quality for each of the nine quality categories in the checklist. Good reports received 75 percent or more of the points possible for each of the PQIs in the relevant quality category, fair reports received 50-74 percent of points possible, and poor reports received less than 50 percent of the points possible. Table 3 shows how each of the nine quality categories performed compared to each other using the above scoring system.

Table 3. Percentage of Reports Rated as Good, Fair, and Poor for Each of the Nine Checklist Categories

Checklist Category	Percentage Good	Percentage Fair	Percentage Poor
Executive summary	29.2	62.5	8.3
Report structure	55.6	31.9	12.5
Objectives & audiences	45.8	51.4	2.8
EQs	27.8	45.8	26.4
Methodology	63.9	20.8	15.3
Findings	6.9	44.4	48.6
Conclusions	63.9	5.6	30.6
Recommendations	76.1	18.3	5.6
Annexes	47.2	22.2	30.6

As seen in Table 3, four of the nine quality categories—recommendations, methodology, conclusions, and report structure—received a good rating in one-half or more of reports, and among these, only the recommendations category received a good rating in around two-thirds of reports. The quality category annexes received a good rating in a plurality of reports. By comparison, two quality categories (executive summary and objectives and audience) received a fair rating in a majority of reports, and one quality

category (EQs) received a fair rating in a plurality of reports. Finally, one quality category (findings) received a poor rating in a plurality of reports, while two quality categories (conclusions and annexes) received a poor rating in more than 30 percent of reports, and one quality category (EQs) received a poor rating in a quarter of reports.

Conclusion 2
The DoS evaluation reports have performed relatively well in their presentation of the evaluation conclusions and recommendations, describing the evaluation methodology, organizing reports, and describing the evaluation objectives. Evaluation reports have performed less well providing complete executive summaries, presenting EQs, and including a complete set of annexes. Finally, evaluation reports have performed relatively poorly presenting a comprehensive set of findings that consider alternative explanations, exploring program effects on different social groups, and describing key evaluation audiences.

4.1.3 Evaluation Report Quality by PQI

Table 4 lists each of the 29 PQIs in the meta-evaluation checklist in descending order of quality, as rated by the coding team. Similar to the overall meta-evaluation score, PQIs that were judged to satisfy the relevant quality criteria in 75 percent or more of reports were rated as good, PQIs judged to satisfy the relevant quality criteria in 50-75 percent of reports were rated as fair, and PQIs judged to satisfy the relevant quality criteria in less than 50 percent of reports were rated as poor. Using this classification scheme, 12 of the 29 PQIs (41.3 percent) were rated as good, 10 (34.5 percent) were rated as fair, and seven (24.1 percent) were rated as poor. Thus, while the plurality of PQIs were rated as good, more than one half of the PQIs were rated as either fair or poor.

Table 4. Ranking and Categorization of Checklist Principal Quality Indicators

Indicator #	Quality Indicator	Number	Percentage
Good			
12	Report describes methodology	72	100
17	Findings draw on all data collection methods	72	100
7	Report describes objectives	70	97.2
5	Report is readable	67	93.0
1	Report has executive summary	66	91.7
11	Report answers all EQs	61	84.7
15	Report describes methodological limitations	59	81.9
27	Annexes include research instruments	58	80.6
22	Recommendations supported by findings and conclusions	56	77.8
13	Report discloses survey and interview sample sizes	56	77.8
26	Annexes include bibliography	56	77.8
23	Recommendations are specific	56	77.7
Fair			
20	Report distinguishes conclusions from findings	52	72.2
25	Annexes include SOW	49	68.1
28	Annexes include list of persons interviewed	47	65.3
6	Report is organized around EQs	46	63.9
21	Conclusions follow from findings	46	63.9
4	Report includes five basic components**	44	61.1
9	Report has no more than six EQs	43	59.7
14	Report describes sampling methodology	43	59.7
16	Report describes data analysis methods	43	59.7
19	Report discusses gender effects	41	57.4
Poor			

Indicator #	Quality Indicator	Number	Percentage
8	Report describes evaluation audiences	33	45.8
24	Recommendations are actionable	33	45.6
10	EQs are that same EQs as in the SOW	32	44.4
3	Report is between 25-35 pages	26	36.1
2	Executive summary includes five basic components**	23	31.9
29	Report includes conflict of interest declarations	17	23.6
18	Report considers alternative explanations	7	9.7

**Project description, EQs, findings, conclusions, and recommendations.

The discussion below brings together the results in Tables 3 and 4 to summarize what role individual PQI scores played in contributing to the quality scores for each of the nine quality categories.

Executive Summary: The executive summaries included in the evaluation reports were, on balance, fair. Indeed, the executive summary was one of the lowest-rated quality categories owing largely to the omission of one or more of the following five key report components: project description, EQs, findings, conclusions, and recommendations. Of the 61 reports (91.7 percent) that included an executive summary, 30 (49.2 percent) omitted the conclusions, and 13 (21.3 percent) either omitted the findings or omitted the EQs. With an average length of six pages, the length of the executive summaries appears reasonable, particularly in the absence of specific guidance from F on the executive summary's length.

Report Structure: The structure of the reviewed evaluation reports was, on balance, good. For the most part, evaluation reports were organized around the primary EQs, addressed the key EQs in the report narrative, and included the five key basic report components, although gaps existed in all cases. Where omissions of basic components did occur, they were concentrated either in the omission of a set of clearly differentiated conclusions in 20 reports (27.8 percent) or by the omission of EQs in eight reports (11.1 percent).

While some key informants complained about overly long, inaccessible, or jargon-heavy reports, these quality indicators did not emerge as significant issues in the meta-evaluation. The meta-evaluation found that the evaluation reports were overwhelmingly readable, meaning that they were written at a level appropriate for non-technical/expert readers and without an excess of technical jargon. While only a little more than one-third of evaluation reports were between the recommended 25-35 pages, the average report length of 38 pages did not exceed the recommended report length by enough to indicate a serious problem relative to the DoS guidelines. (Whether 38 pages are too many in absolute terms is another issue. Some reports may comfortably require fewer than 38 pages, but other reports, particularly those evaluating large or complex programs and/or that cover more EQs, may have a more difficult time covering everything in 38 pages or less.)

Evaluation Objectives and Audiences: Overall, the reviewed evaluation reports did a fair job describing the evaluation objectives and audiences. However, this conclusion masks a significant degree of variation in which nearly all the reviewed reports included a discussion of the evaluation objectives compared to less than one-half of reports that included a discussion of the evaluation audiences.

Evaluation Questions: The evaluation reports did a fair job in presenting the EQs. About 60 percent of reports included the recommended six or fewer EQs, while the actual number of EQs per report averaged only 5.9, barely less than that recommended in the guidelines. Breaking this down further, 30 (41.7 percent) of the 72 reports had between 2-4 EQs, 19 (26.4 percent) had between 5-6 EQs, 18 (25.0 percent) had between 7-9 EQs, and five (6.9 percent) had 10 or more EQs. In addition to the primary EQs, 20 (27.8 percent) of the evaluation reports included a set of sub-evaluation questions (sub-EQs) for an average of 3.4 sub-EQs per report.

It could be argued that the number of EQs is as much a measure of SOW quality as report quality. The sponsoring bureau presumably determines the number of EQs, which is then reflected in the evaluation SOW. In practice, however, evaluators often exercise discretion to increase or decrease the number of EQs, with or without the consent of the sponsoring bureau, although where they do so, it is reasonable to require that they provide an explanation for the discrepancy in the report.

Less than one-half of the EQs in the reviewed reports matched the EQs in the SOW. However, this is largely because, in many cases, the SOW was not included in the report annex, preventing the ET from determining whether the EQs match. Therefore, if we count only those 49 reports that included the evaluation SOW in the report annexes, the EQs in 32 (65.3 percent) of these reports matched the EQs found in the SOW. Among the 17 reports in which the EQs did not match the EQs in the SOW, six reports (35.3 percent) provided an explanation for the discrepancy.

Finally, the reviewed evaluation reports did a good job answering each of the EQs specified in the report in the sense that the reports included a discussion of the relevant findings related to each EQ.

Methodology: The evaluation reports did a good job presenting and describing the evaluation methodology. One hundred (100) percent of the reports included a formal description of the evaluation methodology, over 80 percent described the methodological limitations, and more than three-quarters of reports disclosed the sample sizes of persons surveyed (quantitative) or interviewed (qualitative). The reports, however, did only a fair job of describing the quantitative (44.2 percent of reports) and qualitative (54.4 percent) sampling methods as well as the quantitative (62.2 percent) and qualitative (57.4 percent) data analysis methods.

Findings: The reports did a poor job presenting the evaluation findings for each of the EQs, with only a handful of good reports, making it the lowest-rated quality category in the meta-evaluation. As with the EQs, however, this finding masks a significant degree of variation among the PQIs measuring the quality of evaluation findings. On one hand, the findings in the reviewed reports uniformly drew on all the data collection methods described in the report. On the other hand, fewer than 10 percent of reports included any discussion of possible alternative explanations and only half of the reports included a discussion of potential gender effects.

The 2016 DoS Evaluation Guidance also specifies that bureaus should pay attention to the program's effects on different social groups, in addition to their effect on women. Notwithstanding this guidance, only one-half of the reports included a discussion of the programs' potential effects on different social groups.

Conclusions: Almost two-thirds of reports did a good job presenting the evaluation conclusions, although nearly one-third of the reports did a poor job. Notably, 20 of the 72 reports (27.8 percent) either did not include any discussion of conclusions in the report or did not clearly distinguish the conclusions from the findings, which helps explain why the conclusions followed from the findings in only 63.9 percent of reports. If, however, only those reports that included a set of clearly distinguished conclusions are counted, then conclusions followed from the findings in 88.5 percent (46 of 52) of the reports.

Recommendations: The evaluation reports did a good job including recommendations that followed from the findings and conclusions and providing specific recommendations, but they did a poor job providing actionable recommendations; less than one-half of the reviewed reports provided actionable recommendations. The online survey, KIs, and FGDs confirmed the meta-evaluation findings with regards to the actionability of report recommendations. When discussing the perceived weaknesses of existing evaluation reports, perhaps the most common weakness mentioned by key informants was the presence of irrelevant and unactionable recommendations. As one key informant reported, *"A consistent problem is that the findings and recommendations are superficial and not at all useful, and program officers have*

said they could have learned more just from speaking to grantees. The budgets for these evaluations were \$300-\$400 thousand and up to a year, so it's not that they were underfunded."

Annexes: The evaluation reports did a fair to good job including important information in the report annexes. Approximately two-thirds or more of reports included the evaluation SOW, a bibliography of documents read, copies of the data collection instruments, and a list of persons interviewed. The only area in which reports performed poorly was the inclusion of conflict of interest declarations, which were included in less than one-quarter of reports.

Conclusion 3

DoS evaluation reports have done a good job including a discussion of the evaluation methodology, drawing on all data collection methods in reporting the evaluation findings, describing the evaluation objectives, writing reports that are readable and accessible to lay audiences, and including an executive summary. They have also performed relatively well in terms of describing the evaluation's methodological limitations, including copies of research instruments, disclosing the quantitative and qualitative sample size, providing a set of specific recommendations, and citing important documents read in a bibliography.

In contrast, DoS evaluation reports have done a poor job describing the evaluation audiences, including EQs that match the SOW EQs (in large part due to the omission of a SOW in the report), keeping the evaluation report within the DoS recommended length, providing a set of actionable recommendations, reporting potential conflicts of interest, and considering alternative explanations. Even though the majority of reports included an executive summary, the reports have done a poor job covering essential report components in the summary.

On all other quality indicators, the DoS evaluation results have done a fair job. Overall, the reports' performance on different quality indicators suggest moderate to substantial room for improvement in most quality indicators used for this meta-evaluation.

Overall, the evidence shows that within each evaluation report quality category, there exists moderate to significant variation across individual measures of report quality. In practical terms, this suggests that bureaus require additional capacity development across the board.

4.1.4 Evaluation Report Quality by Report Characteristic

The ET conducted tests to determine whether the average meta-evaluation scores—including the overall scores, quality category scores, and individual PQI scores—differed significantly between coders, year in which the evaluation was conducted, type of bureau (functional vs. regional), type of evaluation (e.g., performance evaluation, impact evaluation, etc.), number of countries covered, and regions. This analysis yielded two statistically significant test coefficients, indicating that functional bureaus earned a significantly higher overall meta-evaluation score than regional bureaus (20.5 vs. 17.4), and reports covering a single country achieved a significantly higher average overall score than reports covering multiple countries (20.8 vs. 17.4). Notably, the test statistic for the report coder was statistically insignificant in all cases, indicating a high level of inter-rater reliability among the coding team during the report scoring process.

A theme that emerged in the KIs is that, relative to functional bureaus, regional bureaus tend (with exceptions) to have lower staffing, less well-trained BECs, and less funding for evaluations, and they tend to support programs whose results are harder to measure. At the same time, functional bureaus have a relatively longer history of conducting evaluations and, according to key informants, tend to be more adept at using evaluations. This discrepancy, to the extent true, might help explain the lower report quality scores earned on average by regional bureaus compared to functional bureaus.

Conclusion 4
Different report characteristics did not, for the most part, explain the reports' relative quality, either overall, by quality category, or by individual PQIs. Notwithstanding, the relative disadvantages possessed by regional bureaus in terms of staffing, funding, capacity, and types of programs supported appear to have contributed to a lower overall quality of evaluation reports compared to reports produced by functional bureaus.

4.2 EQ 2: To what degree are foreign assistance evaluation findings and recommendations being used by bureaus, and what factors contribute to their use or non-use?

Overall Findings
<p>While stakeholders strongly support the importance of doing and using evaluations in theory, actual support, as evidenced by supportive actions, varies widely. DoS bureaus and independent offices only use evaluation findings and recommendations to “some extent” and for a variety of purposes, including to improve program implementation, contribute to bureau learning, inform program design and strategy, support change management, and question basic assumptions. The use of evaluations with bureaus and offices is, on the whole, more sporadic than systemic.</p> <p>The primary factors hindering evaluation use include a lack of leadership support, infeasible recommendations, limited funding, low interest in evaluations, inadequate dissemination of evaluation results, failure to tie evaluations to bureau strategy, fear that evaluations will be used for punitive purposes, absence of a formal knowledge management system with the Department, and poor quality of the evaluation designs as reflected, for example, in their SOWs, data collection methodologies, work plans, understanding of project context, or technical grasp of the relevant subject matter.</p> <p>Evaluation resources and TA provided by F to bureaus and offices have provided critical support in helping certain bureaus and offices move up the evaluation learning curve.</p>

4.2.1 Perceived Importance of Evaluating Bureau Programs

According to over 90 percent of survey respondents, it is either “very important” (83.1 percent) or “somewhat important” (12.3 percent) to evaluate their bureau’s programs. This is not a surprising finding given the makeup of the survey sample in which 45.5 percent of survey respondents were their bureau’s BEC, and 74.2 percent of respondents either had evaluation as part of their job responsibilities and/or had received formal evaluation training. Among this same group of respondents, 52.3 percent were “very familiar” and another 36.9 percent were “somewhat familiar” with the evaluations that had been conducted by their bureau, for a total of 89.2 percent.

The above findings possibly overstate the actual importance that bureaus place on evaluations given the nature of the sample (e.g., largely DoS staff with evaluation responsibility and at least some evaluation experience). Indeed, the KIIs provided a more nuanced perspective on this issue. While KII respondents (drawn largely from the same population as the online survey) strongly agreed that evaluating bureau programs was important, they also made it clear that this perspective was not universally shared by program staff or bureau leadership and there continued to exist resistance to evaluations stretching from program managers up to senior bureau leadership.

The ET found no instances in which resistance to evaluations had translated into outright opposition, at least among those key informants interviewed—everyone the ET spoke with supported at least the concept of doing evaluations—but, as the key informants emphasized, resistance to evaluations manifests itself in a variety of forms, such as the failure to allocate time or resources to support evaluations or pro-forma compliance. As one BEC noted, *“Some bureaus see evaluations as a good resource for learning about programs, and some see them as a checkbox.”*

Resistance to evaluations in turn stems from a variety of factors. Factors noted by key informants included the perceived low priority of evaluations, particularly given the crush of other work responsibilities, political considerations, fear of poor evaluation findings, or a lack of knowledge about evaluation that translates into misunderstandings about its purpose, benefits, costs, etc. (These factors largely overlap with the factors hindering evaluation use and are thus discussed at length in Section 4.2.3.)

Notwithstanding lingering resistance to doing evaluations noted by the key informants, the same key informants also widely acknowledged that the Evaluation Policy has increased the importance placed on evaluations within the Department. This sentiment was strengthened by FATAA, which established legal requirements for bureaus and offices with regards to evaluations and managing for results (MfR). The GAO and the Office of the Inspector General (OIG) have, moreover, communicated an intention to begin auditing bureau and office compliance with FATAA beginning in 2019. Until then, however, F is dependent on the moral persuasion created by the expanded Evaluation Policy and FATAA to encourage compliance.

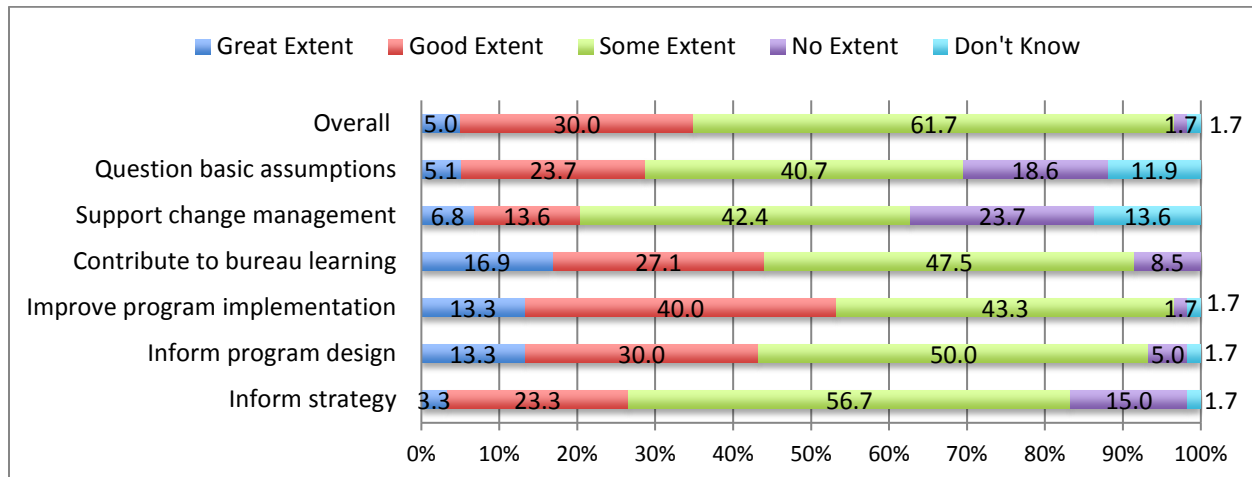
With that said, the ET encountered no examples during the KIIs or FGDs of a bureau or office that has chosen to ignore the Evaluation Policy or otherwise make no attempt to implement it. Each bureau and office visited had in place a BEC and each was, according to the key informants, making an effort to put themselves in compliance with the Policy. While the strength and commitment of this effort varied widely, it is notable, nonetheless, that the effort uniformly existed among the sample of bureaus and offices visited by the ET. The point made time and again during the KIIs and FGDs is that the Evaluation Policy (and later FATAA) put evaluations squarely on the agenda and provided an impetus to doing evaluations that was previously lacking. As one key informant noted, *“Before the Evaluation Policy, we only did a couple of external evaluations. Now, we do two external evaluations per year, and there is more emphasis on evaluation use.”* Another KII reported, *“The Policy helped to solidify the point that evaluations are not going away and that bureaus need to do them.”*

Conclusion 5
<p>Staff at DoS bureaus and independent offices working on or with responsibility for evaluations overwhelmingly believe that it is very important to evaluate bureau and office programs. The theoretical importance attached to doing evaluations, however, is not matched by their practical importance, as measured by actual actions.</p> <p>The Evaluation Policy appears to have been effective at the margin in terms of putting evaluation on the Department’s agenda and motivating improved evaluation practice across a reasonably broad cross-section of bureaus and independent offices compared to the situation beforehand. Although the absence of a counterfactual limits the ability to attribute such changes to the Evaluation Policy, the available evidence suggests that such changes have in fact occurred (although their size cannot be measured in the current evaluation framework), and that the Evaluation Policy has made an important contribution to their occurrence via its effect on the perceptions and motivations of leadership and program staff. In other words, the Evaluation Policy has helped establish a framework for institutionalizing evaluation practice within the Department that was previously absent.</p>

4.2.2 How Bureaus Have Used Evaluations

The survey asked respondents to indicate the extent to which their bureau had used evaluations to inform bureau strategy; inform program design; improve program implementation; contribute to bureau learning; support bureau change management; and question basic assumptions about bureau policies or practices. Their responses are shown in Figure 3, below.

Figure 3. Extent to Which the Bureaus Have Used Evaluations for Different Purposes (N=60)



Less than one-half of respondents said that their bureau had used evaluations for the above purposes either to a “great” or “good” extent, except using evaluations to improve program implementation. In each case, moreover, the majority of responses indicated that the bureaus used evaluations for the relevant purpose only to “some extent.” At the other end of the scale, the most notable response is that around 20 percent of respondents said their bureau used evaluations to “no” extent to support change management within the bureau or to question basic assumptions of the bureau’s policies or practices.

Looking at the broader picture in Figure 3, 61.7 percent of respondents said that, overall, their bureau was using evaluations to some extent, and only 5 percent said that their bureau was using evaluations extensively. An additional 30 percent said that their bureau was using evaluations to a good extent.

KII and FGD participants were also asked how their bureaus had used evaluations. About half reported their bureau is using evaluations to a great or some extent, with most also reporting that their bureau’s use of evaluations has increased since 2012. Below are the primary ways in which key informants said that their bureaus are using evaluations.

Dissemination of Results and Informing Program Learning and Improvement

Nine key informants reported using evaluations to disseminate results to program officers for program learning and improvement, including making mid-program corrections. The following quote is illustrative: *“A recent mid-term evaluation showed that a clearinghouse model did not have the support of stakeholders and was not functioning well. The evaluation found that there was not enough interest in the clearinghouse and there was too little success to do more of it. Thus, after the evaluation, we pushed back on any proposals that would include similar clearinghouse models.”*

One BEC reported that, in his bureau, he sits with the program team to discuss the feedback and whether they concur with the feedback and to what extent:

“The bureau takes the recommendations seriously. The teams do not require the monitoring and evaluation team to follow up too much, but we do it voluntarily. Then we follow up with the program teams to make sure the recommendations are actually being implemented. We also create a two-page informational memo that goes up the chain to the Secretary with recommendations on how we plan to implement them.”

A couple of other BECs also described implementing a recommendation tracking tool to systematically work through how the bureau would respond to the evaluation.⁵ As one described it, *“We have a recommendations follow-up tracking system. First, we sit with teams to discuss the recommendations and determine with them the extent to which each recommendation is accepted and applicable. We prioritize the recommendations that we find most relevant and easy to apply. We then assign personnel to follow up with recommendations selected for use and establish a timeline for implementation.”* A key informant at another bureau that is implementing a similar tool reported that, *“If they [program officers] are not implementing the recommendations, they expect to receive a written justification for why not. The M&E team actually submits the information to the front office so if programs choose not to implement recommendations or brush them off, the information is transmitted to bureau leadership, which gives us leverage to ensure higher rates of compliance.”*

Yet another BEC noted that her bureau has results dissemination sessions at the end of evaluations, circulates results and lessons learned to program teams, and communicates findings to program officers in person. Advocating for the utility of evaluations, another key informant asserted that evaluations *“help address what needs to change to meet the objectives of the program or project and identify if things are going well.”* As a final example, one BEC reported that his bureau is doing an anthology covering the last six years of evaluations the bureau has done, including a four-page summary of each evaluation, that he *“will share with the entire bureau to generate greater support for evaluation and the sharing of lessons learned, pitfalls to avoid, and so forth.”*

Bureau Strategic Planning

Five key informants reported that their bureaus use evaluations to improve bureau strategy and planning, including staffing and funding decisions. As one interviewee noted, *“We use evaluation reports to improve the bureau’s strategy. One evaluation can inform the overall programming approach of the bureau.”* Another BEC reported, *“We have a year-long strategy-making process and use recommendations to make strategies. I engage with program officers to share relevant information from research and evaluations and findings from successful projects we have monitored, and we share bureau-wide lessons learned from that.”* Two other key informants reported that evaluations are used during joint regional strategy reviews.

Program Adaptation

Seven key informants mentioned using evaluation results to make program mid-term corrections. One key informant reported, *“We use evaluation findings and recommendations to make mid-term corrections, especially for improving our program performance management indicators.”* In another case, a BEC’s bureau did a mid-term evaluation of a major program and found that the program did not work. The evaluation helped the bureau understand how to improve its future programming and better understand their stakeholders. Five key informants said that they have taken the results from the evaluation of one program and applied the lessons learned to other programs, while three key informants also mentioned that they used evaluation results to improve their bureau’s overall strategic planning and programming. In at least one case, a key informant’s bureau shut down a program because an evaluation found it to be ineffective.

⁵ The Evaluation Policy specifies that, “Once the evaluation is completed, bureaus and independent offices should respond to evaluation recommendations with a written summary to bureau or independent office leadership” (p. 7). F has developed a recommendation tracking tool to help bureaus comply with this requirement. However, it was not evident from the interviews whether the tracking tool used by these bureaus was the F tool, a tool that they borrowed/adapted from another source, or a tool of their own creation.

Grants Management and M&E System Improvement

Two KIIs mentioned that evaluation results are used to improve the grants management process. For example, a BEC reported that her bureau works with smaller local organizations overseas, which *“although best situated on the ground, are not always best qualified to implement.”* Based on evaluation results, her bureau shifted to implementing small grants by channeling money to larger grantees that, in turn, give sub-grants to the smaller organizations and provide them with capacity development. The same BEC reported that her bureau has used evaluation recommendations to speed up their report submission and review process. Another key informant reported that evaluations results led his bureau to require grantees to submit an M&E Strategy Paper and have more robust M&E systems.

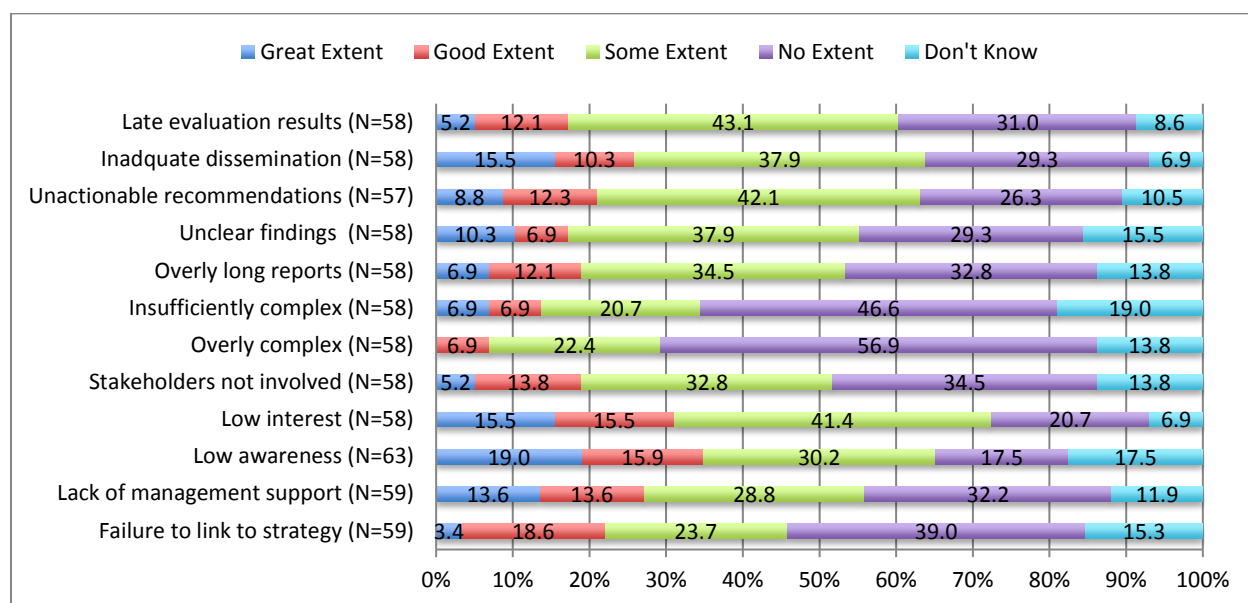
Conclusion 6

Evaluation use in the Department is inconsistent and yet to be institutionalized across bureaus and independent offices, albeit with some exceptions. Where evaluations are used, they are used for a variety of purposes, with the most common purposes being to inform program design and improve program implementation.

4.2.3 Factors that Hinder Bureau Evaluation Use

Given that bureaus are only using evaluations to some extent, the survey posed two questions, one prompted and one unprompted, asking respondents to identify which factors most hindered evaluation use. In response to prompting (providing respondents with a list of possible hindrances and asking them to rate their importance), respondents identified the following factors as the strongest hindrances to evaluation use: low awareness about evaluations; low interest in evaluations; inadequate dissemination of evaluation results; and lack of management support for evaluations (Figure 4). Less important, but also relatively frequently noted hindrances included the failure to link the evaluation to bureau or regional strategy as well as recommendations.

Figure 4. Factors Hindering Bureau Evaluation Use (Prompted)



What is notable in Figure 4 is that respondents demonstrated reluctance to say that any of the factors hindered evaluations either to a “great” or “good” extent; rather responses tend more to be clustered in the “some extent” or “no extent” categories. This finding suggests a lack of consensus among respondents as to which factors are hindering evaluation use, together with the perception that no factor or set of factors predominate in hindering evaluation use.

When asked unprompted which were the most important factors hindering evaluation use, the responses identified low leadership buy-in as the primary hindering factor, followed by non-actionable or irrelevant recommendations, lack of funding, and evaluations that are not tied to bureau strategy (Table 5). The results in Table 5 do not indicate that one hindrance or set of hindrances dominates. For example, low leadership buy-in is the most frequently cited hindrance, but it was only cited as the most important hindrance by 11 (15.3 percent) of respondents, and as a top-three hindrance by 17 (23.6 percent) of respondents. Although the ordering of importance in Table 5 does not track precisely onto Figure 4, it confirms the broader finding that respondents do not agree on which are the most important factors hindering evaluation use.

Table 5. Factors Hindering Bureau Evaluation Use (Unprompted)

Factor	1 st Most Important	2 nd Most Important	3 rd Most Importance	Total
Low leadership buy-in	11	6	1	17
Non-actionable or irrelevant recommendations	4	8	3	14
Lack of funding	8	1	1	10
Evaluations are not tied to bureau strategy	5	2	2	9
Evaluation not well designed/executed	3	1	2	6
Poor dissemination of evaluation results	4	1	1	6
Reports not concise or readable	5	1	0	6
Evaluation results are not timely	3	0	1	4
Low leadership/staff knowledge	0	3	1	4
Low program management capacity	2	0	1	3
Low stakeholder involvement	1	2	0	3
Bureaucracy	1	1	1	3
Misunderstanding of evaluation purpose	0	2	1	3

KII and FGD respondents cited several factors that they believe have hindered evaluation use within their bureau or office, many, but not all, of which correspond to the factors covered by the online survey. The factors mentioned most frequently in the KIIs and FGDs are discussed below. (The same factors largely help explain the perceived importance of doing evaluations, as was noted in Section 4.2.1.)

Lack of Leadership Buy-In and Engagement

Nine key informants reported that the lack of leadership buy-in and engagement is a major hindrance to evaluation use, particularly given that buy-in from leadership is key to facilitating data-driven decision-making, including decisions about funding allocation. As one key informant put it, *“Without front office buy-in, the program office is often left without additional money to implement good recommendations.”* Some senior bureau leaders do encourage the use of evaluation results and follow-up on whether evaluation report findings and recommendations are used, as this BEC described: *“When you operate in a leadership-supported environment, it is also easy for us to present the findings and recommendations we receive from evaluation reports to the leadership.”* These bureau leaders, however, appear to be much more the exception than the rule.

F is aware of this issue and is making efforts to address it. It has, for example, created a PowerPoint presentation called “10 Things to Know about Evaluation at State” as a resource for BECs and other bureau evaluation staff when they talk to senior bureau leadership. The presentation, which F has also presented at fora attended by bureau leadership, addresses 10 key questions or concerns people have about evaluation that F has *“heard over and over since 2012.”* F is also in the process of working with the Foreign

Service Institute (FSI) to integrate an evaluation module in the training curriculum for ambassadors and other higher-level officials posted to overseas assignments.

Lack of Understanding and Awareness of Evaluation

Three KIIs found that a lack of understanding and awareness of the purpose of evaluations hinders their use. One key informant reported, *“At the higher levels, there is a lack of understanding about the benefits of evaluation and little appreciation of why it is important.”* Key informants referred frequently to the potential benefits of doing evaluations—including program design and adaptation, institutional learning, informing programming decisions, and enhancing public accountability—and they expressed concern that such potential benefits were poorly understood or appreciated among Department leadership and program officers. If these purposes and benefits were better understood, according to this line of reasoning, then this would contribute to reducing resistance to doing evaluations and using their results.

Fear of Adverse Consequences

One commonly cited result of the lack of understanding discussed above was the fear that evaluations were akin to an audit. Even if their purpose was not punitive, per se, the fear is that they will be used for that purpose. One key informant, for example, recounted the story of a program manager in another bureau who was *“so angry about negative findings that he wanted to pull funding from the evaluation contractor to make sure the final report never saw the light of day.”* In many cases, he said, *“leadership is terrified of reports going public because they are worried about evaluation being used as a punitive tool versus using it as a learning tool.”* Thus, the perception that *“if there are negative findings, then there will be budget cuts or other adverse repercussions”* is common across the Department, which creates a natural resistance to doing evaluations.

Political Considerations

Many bureaus work in politically sensitive areas, which require a substantial degree of discretion and the management of sensitive relationships. One interviewee explained that evaluations, and using them as the basis for programming decisions, can risk calling attention to the relevant program when the bureau would prefer the program to remain more under the radar.

Political considerations, and the fear of adverse consequences, can also motivate bureaus to refrain from posting evaluation reports to the evaluation registry. For example, one key informant recounted how a bureau balked at submitting a report even after it had been requested by the Congressional Research Service owing to a desire to avoid calling attention to the program to the host country government and in the end submitted a heavily redacted report to Congress.

Lack of Bureau Capacity or Resources

KIIs revealed that the ability of bureaus to incorporate evaluation results into programs depends on the capacity and resources that individual bureaus have to support the process. Many key informants agreed, for example, that bureaus that are better-funded and have experienced M&E staff are much more likely to use evaluation results than bureaus that lack adequately trained staff or appropriate funding for monitoring, evaluation, and learning.

Capacity constraints become particularly acute when a bureau supports programs whose results are less amenable to direct measurement or harder to attribute to program interventions. As one key informant noted,

“Many of the things that the State Department and diplomats try to do are difficult to measure, so you get evaluations that focus on areas with more easily collected metrics. Also, sometimes you have a metric that works, but it is difficult to attribute what part of the result comes from the

bureau's intervention. These issues can lead to evaluations that are not aligned with what bureau leadership sees as the most important things they are doing and want evaluated."

Key informants further indicated that bureaus with a relatively long history of conducting evaluations are better funded and staffed and, therefore, more prepared to use the evaluations. One BEC reported, *"Where evaluations have been used more, program officers have been involved in every step of the process and can articulate what they want, create good evaluation questions, are pragmatic about what they receive, manage trade-offs effectively, and are realistic but pushy."* Another key informant explained, *"The M&E team in my office promotes doing evaluations, but the program teams are the ones that have to do it."* She noted the importance of having program managers with the technical capacity to use evaluation findings and enough time to spend on doing it.

Poor Quality Evaluation Designs and SOWs That Lead to Poor Evaluations

Four KIIs reported that limited bureau capacity and program officer engagement is a major factor in producing poor quality evaluation designs and SOWs, which tend to lead to poor evaluations with irrelevant and unactionable recommendations. One key informant from a regional bureau reported, *"Prior to the current BEC, people were very disappointed with the results of evaluations due to poor SOWs and not asking the right evaluation questions. Evaluators answered the questions well enough, but they were not the right questions, so the evaluators didn't get the evaluation to where they wanted."* Similarly, another BEC remarked, *"If program officers don't know how to ask the right evaluation questions or ask unanswerable questions, they wind up with a bad evaluation report."* Other key informants reported that SOWs are often not written with a budget or period of performance that allows evaluators to do in-depth, quality evaluation reports.

Findings and Recommendations That Are Not Relevant or Actionable

Six key informants reported that many recommendations are not relevant or actionable. According to one key informant, for example,

"Recommendations are often not grounded in reality or actionable. They might be great ideas, but you simply can't implement them. For example, one evaluation dealt with a project that someone manages with 30 to 40 percent of his time. However, the evaluation recommendations seem to assume that you needed multiple people to spend a great deal of time implementing the recommendations. It is important to connect recommendations to current staffing and resource levels."

Similarly, a BEC noted, *"Any recommendations that would require extra staff, funding, systems, or other additional resources are difficult to implement, so we ask that evaluators not make recommendations of that sort."* The same BEC reported that evaluations often include recommendations that are not relevant to her bureau, so she asks evaluators to target their recommendations to specific groups like the bureau, government, or international organizations, because they *"do not have control over implementation of any recommendations outside of the bureau."*

Five FGD participants also reported that there is a big gap between the micro-level data that come in from the program-level evaluations and information required to make programming decisions at the macro- or policy-level. *"It makes me skeptical when others say they will make decisions based on data-driven results,"* noted one FGD participant, *"because the data simply isn't there to inform those decisions."*

Lack of Cross-Bureau Knowledge Management or Mandate

Focus group participants observed that, often, there are no department-wide mechanisms for helping ensure bureaus incorporate evaluation findings and recommendations for program learning and

improvement. While such mechanisms may exist in individual bureaus or offices, these are the exceptions. One key informant elaborated at length on this issue:

“When you try to grow a body of knowledge, you need someone who says to the team, ‘Make sure you read the evaluations.’ You need someone responsible for pushing that information to the right people; you have to take the knowledge to the user rather than expecting people to take the knowledge out. You need someone whose job it is to connect lessons between Kenya and Southeast Asia, for example, by facilitating learning, identifying opportunities for sharing information, etc. Evaluations are not a library but a book mobile. Evaluation is important, but there needs to be somebody whose day job is to connect the dots and facilitate learning and identify opportunities for sharing information. We do this in our bureau and it is what makes us unique. So, we need knowledge management, evaluation, and learning to be thought of being together.”

In addition, some key informants reported that there is no cross-bureau enforcement mechanism for the use of findings and recommendations. According to one,

“The lack of a real mandate contributes to the non-use of evaluation findings. People are not being penalized if they don’t go through the process of tracking and managing evaluation recommendations. There is a lack of momentum in some bureaus because there is no mandate from the Secretary’s office that reinforces the benefits of evaluations and potential for program learning and improvement.”

Missing Data in the Evaluation Report

Another factor identified by five key informants was that evaluations often do not provide a decisive recommendation for implementation due to budget, time, or other constraints under which the evaluators were working. This, in turn, adversely affects the evaluators’ ability to generate useful information. As noted by one key informant, *“Almost every evaluation says that there’s not enough time or data to determine whether there is an impact, even though most programs have been going on for several years and in some cases for decades.”* Alternatively, the evaluation may only cover a slice of a multi-country or otherwise complex program, which limits the evaluators’ ability to conclude anything meaningful about the program as a whole.

Evaluation Reports That Are Difficult to Read or Targeted to the Wrong Audience

Five key informants further mentioned that evaluation reports that are readable and exclude scientific jargon are more likely to be used than otherwise. One of the key informants indicated that he was discouraged by one of the evaluation reports his office sponsored because it was written as if for a scientific audience such that he was not able to translate the results to the front office and get them to look at it. Some key informants also observed that longer reports are less likely to get read or used. They emphasized that a succinctly written report that synthesizes pertinent information in a relatively short narrative, supported by infographics summarizing all key findings and recommendations, will help them distribute the report within the bureau and promote the use of the findings and recommendations among bureau leadership and program officers.

Conclusion 7

Several factors contribute to hindering evaluation use within the Department that are described in the findings; however, there is no consensus as to which of these factors have played the biggest roles in hindering evaluation use.

With regards to knowledge management, F's efforts to increase the dissemination of evaluation results via the evaluation registry and evaluation CoP fall short of creating the systematic knowledge management processes needed. Further, the fear that poor evaluation results will be used for punitive purposes, particularly to reduce program funding, is an important and understandable concern. Any campaign to change the Department's evaluation culture must confront this issue better than what has been done to date.

F's Tools to Address Factors Hindering Evaluation Use

F has created a high number and wide variety of tools to help bureaus address many of the hindrances to evaluation use discussed above. These include tools covering almost every phase related to conducting evaluation, including: preparing SOWs; writing EQs; designing, conducting and managing evaluations; writing evaluation reports; disseminating evaluation results; tracking implementation of evaluation recommendations; managing the post-evaluation management response; and addressing common questions or concerns about doing evaluations. F has also created an evaluation wiki, produces a bi-monthly State of Evaluation newsletter, and helps BP support the DoS' evaluation CoP.

While F has invested considerable effort to publicize the above tools, it does not have a structured approach for advising DoS staff on how to use the tools. For example, the evaluation courses offered jointly by F and BP have not incorporated any of the tools in the courses. Neither has F, to date, provided many examples of how the tools have been used in practice (although it has stated an intention to do this more in the future). Thus, BECs and other DoS staff are largely left to their own devices on deciding which tools to use and on how to apply them.

Discussions with key informants revealed that those BECs at the most advanced end of the spectrum (highest capacity) were less likely to use the F tools, preferring to use their own tools or access them from different sources, while those at the least advanced end of the spectrum (lowest capacity) were more likely to use the tools but often struggled in figuring out how to use them. Key informants among this second group expressed a strong demand for more examples on how to apply the tools.

Conclusion 8

While F has developed a suite of tools addressing the range of factors hindering evaluation use, it has not instituted a process to instruct BECs and other DoS staff on how to use them or provided many practical examples of their use. Thus, BECs and other DoS staff who need these tools the most, and who are the most likely to use them, can find it difficult to use or adapt them.

4.3 EQ 3: To what extent have bureaus met the baseline requirements of the policy for planned and completed evaluations and staffing and utilizing the Bureau Evaluation Coordinator position?

Overall Findings

Each year, from 2014-2016, about 90 percent of bureaus and independent offices complied with the DoS Evaluation Policy by completing at least one evaluation per year or having at least one evaluation ongoing at the end of the year. In terms of provisions beyond the one-per-year requirement, less than half of key DoS staff interviewed believe their respective bureaus and independent offices are

compliant with the Evaluation Policy. Overall, compliance with the Evaluation Policy varies across bureaus and independent offices.

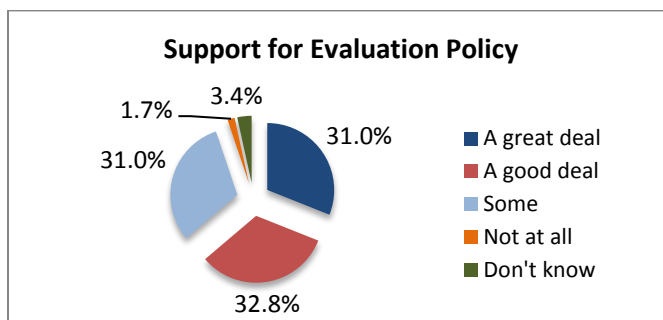
While all bureaus and independent offices have staffed and are utilizing the BEC position, the roles and responsibilities of BECs, the amount of time BECs spend on evaluation activities, and the experience and capacity of BECs vary significantly across DoS bureaus and independent offices. BECs range from highly experienced, visible, well resourced, supported, and proactive to inexperienced, inconspicuous, poorly resourced and supported, and reactive, without enough time or capacity to promote evaluations. The majority of Department evaluation staff independently consult with their BEC on evaluation issues, with most of them finding the consultations useful. However, many do not consult with their BEC or do not find the consultations useful.

The above discrepancies, together with a lack of senior leadership buy-in, constitute primary challenges to achieving department-wide compliance with the Evaluation Policy. Other challenges include a lack of M&E staff and technical expertise among existing staff, funding constraints, programs that do not easily lend themselves to evaluation, and bureaucratic external contracting obstacles. Overall, and owing in large part to the same challenges, evaluations have yet to play an important role in informing bureau policy, programming, and decision-making, and most BECs do not regularly engage bureau leadership, with exceptions in both cases on a case-by-case basis.

4.3.1 Familiarity with and Support for DoS Evaluation Policy

About 95 percent of survey respondents were either “very familiar” (51.7 percent) or “somewhat familiar” (43.1 percent) with the DoS Evaluation Policy (Figure 5), while all key informants indicated that they were familiar at least to some degree with the requirements of the Evaluation Policy.

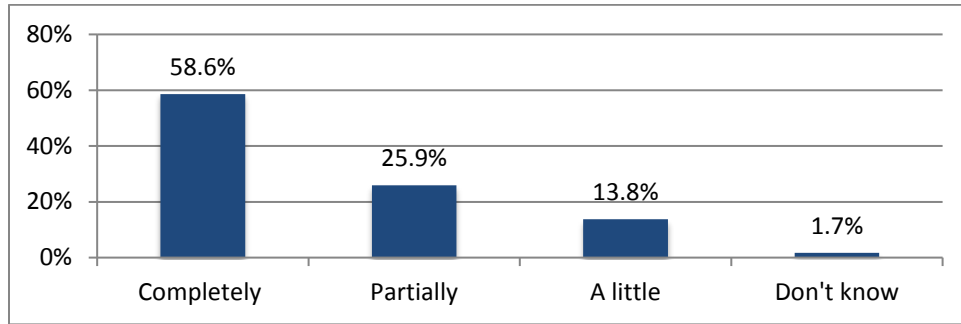
Figure 5. Leadership Support for the DoS Evaluation Policy (N=58)



A majority (58.6 percent) of survey respondents further believe that the DoS Evaluation Policy applies “completely” to their bureau (Figure 6), while more than one-third said that the policy only applies “partially” (25.9 percent) or “some” (13.8 percent) to their bureau (Figure 6).

The results in Figures 5 and 6 should be viewed with some caution, given that the survey was administered to the group of people most likely to be familiar with the details of the Evaluation Policy. This is not universally the case in the Department, however, particularly as one moves up the hierarchical ladder. For example, one senior bureau leader interviewed stated, “A challenge in implementing the evaluation policy is that nobody knows about it...for example, the requirement in the policy that any pilot program has to have an evaluation is something that most people don’t realize is a requirement...when told that the policy requires at a minimum one evaluation per year for each large program, the principal deputy secretary said it was ‘news to him.’” This is a misrepresentation of the Evaluation Policy, which requires that pilot programs be evaluated before being scaled up and that large programs be evaluated once in their lifetime or every five years, not once per year. It is an example that highlights how the Evaluation Policy is still misunderstood among senior leadership.

Figure 6. Extent to which the Evaluation Policy Applies to the Bureau (N=58)



Notably, strong, if not universal, support for the Evaluation Policy existed among key informants interviewed for the evaluation—the majority of whom recognized in KIIs the positive role that the Evaluation Policy has played in motivating increased emphasis on evaluations and improved evaluation practice within the Department.

Conclusion 9

Overall, familiarity with and support of the DoS Evaluation Policy in the Department is relatively high but far from universal in both cases. While F has invested a significant amount of time and effort to educate Department staff and leadership about the Evaluation Policy, significant gaps still remain. Thus, notwithstanding the efforts F has made to this point, there remains further work to do to ensure that information about the Evaluation Policy penetrates to and is internalized by all staff and leadership levels within the Department.

4.3.2 Bureau Compliance with DoS Evaluation Policy Requirements

Bureaus Meeting Evaluation Policy Requirements

Arguably, the key requirement in the DoS Evaluation Policy is that bureaus evaluate a minimum of one program per year. If bureaus do not comply with this requirement, then all other provisions related to the dissemination, use, and tracking of evaluations are rendered moot. In this light, the 2016 Bureau Evaluation Plans show that 16 out of 22 bureaus had completed at least one evaluation that year (Table 6), while four other bureaus had at least one evaluation ongoing at the end of the year. In 2015, 13 out of 22 bureaus completed at least one evaluation, and another seven bureaus had at least one evaluation ongoing at the end of the year. In 2014, 22 out of 25 bureaus completed at least one evaluation, and no bureaus had evaluations ongoing at the end of the year.

Table 6. Compliance with Evaluation Policy Requirement to Conduct at Least One Evaluation per Year

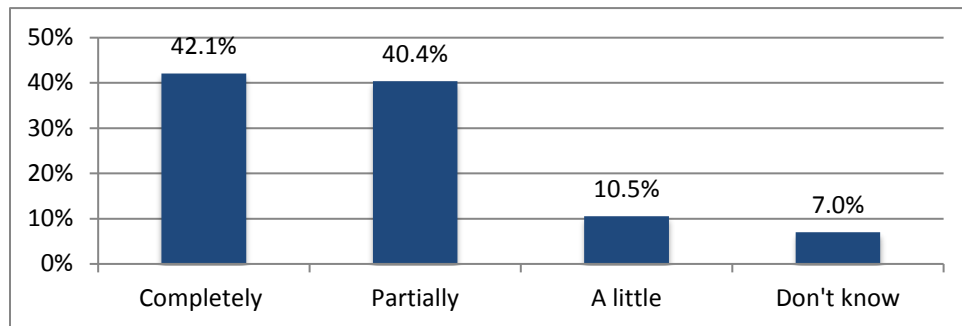
	2014	2015	2016
Number of bureaus/offices	25	22	22
Number of bureaus with completed evaluations	22 (88.0%)	13 (59.1%)	16 (72.7%)
Number of bureaus with ongoing evaluations	11 (44.0%)	15 (68.2%)	16 (72.7%)
Number of bureaus with ongoing but not completed evaluations	0 (0.0%)	7 (31.8%)	4 (18.2%)
Number of bureaus with completed or ongoing evaluations	22 (88.0%)	20 (90.0%)	20 (90.0%)

Conclusion 10

The majority of bureaus and independent offices have complied with the DoS Evaluation Policy to conduct at least one evaluation per year during 2014-2016, although the percentage of bureaus and offices complying with this requirement has fluctuated from year to year and was lower in 2016 than in 2014. If the bureaus and offices that had an evaluation ongoing at the end of the year are counted, then nearly all bureaus and offices were in compliance with the policy in each year from 2014-2016.

When asked in the online survey about the extent to which their bureau or independent office had complied with the Evaluation Policy, 42.1 percent said that the bureau has “completely” complied and another 40.4 percent said that the bureau has “partially” complied (Figure 7).

Figure 7. Extent to Which the Bureau Has Complied With the Evaluation Policy (N=58)



As this survey question did not specify any particular provisions of the Evaluation Policy, the answers to this question appear to indicate the respondents' perceptions of their bureaus' compliance with the overall policy. Indeed, comparing the answers to this survey question with the information from the bureaus' Evaluation Plans presented above supports this interpretation. This interpretation, moreover, jibes with the key informants' responses during the KIs and FGDs in which their answers to a similar question were given within the context of the overall Evaluation Policy, as opposed to specific provisions of the policy.

Below are some examples of how key informants reported the extent to which they believe their bureaus are meeting Evaluation Policy requirements and, accordingly, taking actions to comply. Note that these findings are based on the perceptions of bureau personnel and not on the basis of an actual audit of bureaus' compliance with all provisions of the Evaluation Policy.

Eleven (11) key informants reported that their bureaus are meeting or exceeding the baseline requirements of the Evaluation Policy. One key informant reported that his bureau tries to set aside three to five percent of each program budget for M&E, has done 13 evaluations out of 19 large programs currently being implemented, and evaluates some programs that are not large. Three other key informants reported that their bureaus did five to seven evaluations in the past year, including applied research evaluations, baselines, and quasi-experimental evaluations⁶.

One BEC reported that—before the Evaluation Policy was updated to include performance monitoring requirements—his bureau was already working towards actions that would help the bureau comply. He added, *“Now, for each project, we develop an activity concept paper with templates showing how we are aligned with the policy. We identify performance goals, what resources we need, and timeline—the things that are necessary to do good performance management. We are well-positioned to comply with new regulations.”*

⁶ Baselines are done before or at the start of a program and provide a basis for planning and/or evaluating program results in the future. They do not count as an evaluation under the DoS Evaluation Policy.

Another key informant stated, *"We are ensuring that we are doing evaluations, and we are doing more than is required. We are doing M&E for all of our programs, and we do an annual evaluation for all six programs."* One BEC asserted, *"We are on track to meet the requirements of the Department of State policy on evaluation. The steps we are taking include defining projects and programs, defining large and medium programs, identifying indicators for different programs. We are also briefing other bureaus about FATAA."*

Bureaus Unsure About Meeting or Partially Meeting Evaluation Policy Requirements

Reflecting half of survey respondents who said either their bureau is partially compliant with the Evaluation Policy or that they do not know whether their bureau is in compliance with the policy, 11 key informants reported being in full compliance while 5 reported they are not in compliance and the rest were unsure about their compliance. One key informant, for example, indicated that his bureau was fully compliant with evaluating at least one program each year, but the bureau is still working to comply with the program management provisions in the broader policy encapsulated in 18 FAM 300. He explained,

"We are starting to use program management guides from F. When we do a SOW, we send it to F to review and give feedback. We are sending people on our team to The Evaluation Institute to get training in addition to what they get from F. We are working to keep the evaluation registry updated. 18 FAM 300 makes it very clear that we must do logic models and have theories of change, so we are trying to develop a new approach to standardize this in our M&E practice."

Another key informant reported that his bureau is in compliance with the Evaluation Policy for the most part, with the exception of the dissemination provisions: *"We have outlined all the policies we need to be in compliance with, but one area in which we are lacking still is dissemination. We posted the most recent evaluations on the evaluation registry, we have logic models and an M&E plan, and we are doing at least one evaluation per year. We are looking at a 2- or 3-year horizon for constructing an evaluation strategy."*

One key informant was less sure of her bureau's compliance with the Evaluation Policy, but thinks it is in a good position.

"We have applied for the evaluation funding competition, so this year we will be doing a major evaluation of a program that has been going on for five years. If everything goes as planned, we should be able to do one big evaluation per year. In the absence of the competition, however, we are not sure we would have done an evaluation this year. We applied for and got more money both in support of the evaluation policy and to comply with it, but also because it is difficult for us to shoulder the entire burden of doing evaluations."

According to yet another key informant, *"We are in compliance with the evaluation part, but it will take time for us to be in compliance with the program design and monitoring side [in the expanded 18 FAM 300], as we had not been doing it previously. We will be able to do it, but manpower is a constraint, as it will require more effort to get into compliance."*

Two key informants reported that it is difficult for them to determine whether they are compliant because the Evaluation Policy is not sufficiently clear about what qualifies as a program and, furthermore, what qualifies as a large program (notwithstanding that these are clearly spelled out in the Evaluation Policy along with F's ongoing efforts to communicate these basic policy provisions to the bureaus). To demonstrate this difficulty, one of them noted that, while her bureau complies with the policy, *"It is not as simple as that, because we do not have a nice list of programs based on funding amounts. We receive voluntary contributions to international organizations, so the money is comingled, and we do not know where it goes sometimes."*

Bureaus Not Meeting Evaluation Policy Requirements

Five key informants believe their bureaus are likely not meeting the policy requirements in whole or in part, noting that they are doing one evaluation per year but are not evaluating ongoing programs once every five years or large programs once in their lifetime.

A smaller number of interviewees reported that their bureaus are not meeting the Evaluation Policy requirements. One key informant admitted, *“We are not prepared now, but we look forward to working with F as we try to meet the requirements of the policy.”* Another interviewee reported, *“The bureau is not in compliance with the policy. We have a BEC in name only and a fragmented approach to doing evaluations. While we are attempting to comply, I am not aware of what is happening in the bureau and thus cannot answer what the bureau is doing writ large. I also don’t know of any actions being taken.”*

Several key informants expressed skepticism whether other bureaus are meeting the policy requirements, because the policy lacks any kind of enforcement mechanism. Key informants also expressed the general sentiment that regional bureaus are much less likely to be compliant with the policy than functional bureaus because it is more difficult to apply the policy to diplomatic engagement programs, which are more common in those bureaus.

It is worth noting here that based on data collected and reviewed, the ET believes F has been both very active in helping bureaus understand how the Evaluation Policy applies to them and quite flexible in working with bureaus to help them comply with the policy. Nonetheless, as indicated by the KII responses, there remain lingering questions and confusion about compliance requirements and how the policy applies to individual bureaus or offices.

Conclusion 11

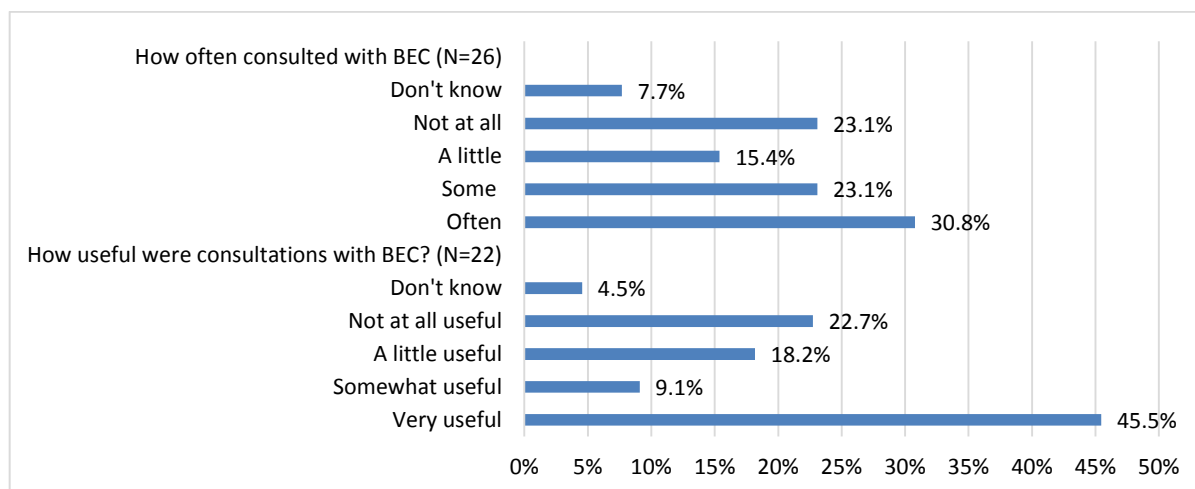
When compliance with the DoS Evaluation Policy includes all provisions of the policy, bureau, and independent office, compliance with the policy falls to less than one-half, according to the perceptions of key bureau and independent office staff. Thus, while bureaus and offices are, or for the most part, doing the minimum number of evaluations required by the Evaluation Policy, they are having a harder time complying with other provisions of the policy. From the context of the findings, their difficulties stem in large part from the MfR requirements in the revised Evaluation Policy.

Many DoS evaluation staff appear to have different views, in addition to some lingering confusions, about what compliance means, notwithstanding language discussing these issues in the Evaluation Policy and F’s considerable efforts to clarify all these points. For example, questions linger about how a bureau operationalizes the requirement to evaluate large programs once in their lifetime when the bureau supports multiple small programs, many of which are funded at relatively low levels, making the benefits of evaluating them questionable relative to the costs. Confusion also appears to exist about what compliance means with regards to the MfR requirements in the revised Evaluation Policy.

4.3.3 Use of BECs and BEC Roles, Responsibilities, and Visibility

Among survey respondents who are not BECs, nearly 70 percent have consulted with their BECs either “often” (30.8 percent), “some” (23.1 percent), or “a little” (15.4 percent) (Figure 8). Of these, about half found the consultations to be either “very useful” (45.5 percent) or “somewhat useful” (9.1 percent), while 40.9 percent found the consultations to be either “a little useful” (18.2 percent) or “not at all useful” (22.7 percent).

Figure 8. Use of and Usefulness of BEC Consultations



Conclusion 12

The majority of Department evaluation staff are consulting with their BECs on evaluation issues and find those consultations useful. At the same time, however, many evaluation staff either are not consulting with their BECs or do not find these consultations useful. Alternative interpretations of this finding are that many non-BEC evaluation staff are not aware of their BECs (less likely) or do not believe it is worthwhile to consult with them (more likely), with the latter possibility supported by the fact that several bureaus lack adequately trained or experienced BECs and that a relatively large share of non-BEC evaluations staff do not find consultations with their BEC to be particularly useful.

KIIs brought to light the wide range in BECs' experiences and roles across bureaus. BECs vary in terms of their prior evaluation experience, their technical capacity, their responsibilities, the amount of time they devote to evaluation activities, their visibility and influence within the bureau, and the level of bureau financial and other support they receive. Not surprisingly, there was also evidence from the KIIs of a more-or-less direct relationship between where the bureau was on this continuum and how far along it was in institutionalizing evaluation practice within the bureau.

Eleven (11) KIIs reported that BEC responsibilities were a full-time job or a major portion of the BEC's overall work responsibilities. In these cases, moreover, the BECs played a role within their bureaus that went beyond evaluations to include responsibility for activities falling generally under M&R, including activities such as performance monitoring; SOW, logframe, theory of change, and indicator development; program design; reviewing M&E plans; data quality assurance; and so forth.

Three key informants reported that their BECs' responsibilities took from 10 to 30 percent of their time. They reported that their role as BEC largely focuses on making sure their bureaus are in compliance with the Evaluation Policy, managing the evaluation contracting process, and advocating for M&E. They each noted their role sometimes includes advocating for evaluations to senior leadership in an effort to change the evaluation culture within their bureaus. An additional five BECs reported spending little-to-no time on evaluation-related tasks and were BECs largely in name only.

The BECs' visibility also varies widely from bureau to bureau. Certain BECs claimed to enjoy a relatively high level of visibility, others a moderate amount, and others very little. One BEC, for example, said that many program officers in her bureau did not even know that BECs exist. The BECs who reported that their bureaus place a higher level of importance on evaluation had greater internal evaluation capacity, and

the BECs whose job duties were more expansive tended to report a closer working relationship with program officers providing a range of MfR services.

One additional detail noted in the KIIs is the lack of a uniform understanding of what the BEC position entails, which a number of key informants mentioned as a likely contributing factor to the wide diversity found in BEC roles and responsibilities. According to one key informant, *“There is no set standard of tasks, no standard department-wide language about what the BEC should do, and no set position description, so roles vary widely.”*⁷

Conclusion 13

The roles and responsibilities of the BEC, and the amount of time he/she spends on evaluation-related activities, vary significantly across DoS bureaus and independent offices. The BECs who are most active and who spend more time on evaluation-related activities also tend to have a wider portfolio of responsibilities that include activities falling broadly under MfR, which consume a significant amount of their time alongside their evaluation-related activities. While the language in the Evaluation Policy and the DoS Evaluation Guidance spell out the minimum requirements and tasks for the BEC, they are fairly general and do not specify/suggest minimum qualifications, thus leaving ample room for discretion or confusion.

4.3.4 BEC and Bureau-Wide Evaluation Capacity

Twenty-four (24) of the 30 (80 percent) BECs who responded to the online survey had received previous evaluation training. Among the same group of 30 respondents, five had no evaluation experience, two had less than a year of evaluation experience, five had one to two years, 11 had between three to five years, five had between 6-10 years, and two had more than 10 years. Of the six BECs who have not received previous evaluation training, one had three to five years of evaluation experience, and five had no evaluation experience. Thus, 5 (20 percent) of the 30 BEC respondents had no evaluation experience, nor have they received any evaluation training, and another 7 BECs (23.3 percent) had fewer than two years of evaluation experience.

The existence of a significant proportion of untrained and/or relatively inexperienced BECs found in the online survey was confirmed in the KIIs. More of the BECs interviewed acknowledged having relatively little evaluation experience than those who reported having relatively significant evaluation experience prior to their appointment as BEC, a phenomenon noted by this senior bureau leader: *“Some bureaus have identified people to be the BEC who really are not the best candidates.”* Another interviewee similarly reported that, because of the Evaluation Policy:

“Suddenly bureaus have to assign a BEC, but that doesn’t mean they necessarily have anyone on staff who is technically capable of taking on the role. So the bureau often turns to someone who is wearing another hat and adds this to their duties. For larger bureaus that are serious about utilizing evaluations, the BEC is expected to evolve into a full-time professional position. For now, in small bureaus, however, it is not realistic to expect full-time expert staff to be assigned to the BEC position.”

One interviewee described herself as the default BEC selected for this role because *“there was nobody else,”* which has been a challenge because she *“doesn’t have the educational background or the*

⁷ The Evaluation Policy describes the BEC roles and responsibilities this way: “Each bureau or independent office must identify a point of contact with decision-making authority to serve as the Bureau Evaluation Coordinator to ensure that the evaluation function is fully operational and integrated into the planning and decision-making process. He or she will serve as the main point of contact in the bureau on evaluation and will coordinate monitoring and evaluation activities and interact with BP and F on the bureau’s compliance with this policy” (p. 5).

experience to function in this capacity effectively.” Some BECs, conversely, reported that they had worked in M&E for many years before being appointed as BECs. The ET, however, also spoke to BECs who were not only highly experienced but were also highly capable, and who performed a wide variety of M&E tasks for their bureaus, encompassing both evaluations and MfR activities.

As for bureau-wide staff evaluation capacity, key informants were evenly split between those reporting high evaluation capacity and those reporting medium or low evaluation capacity. Bureaus with less evaluation capacity also tended to have BECs with less evaluation experience and knowledge and vice versa. Bureaus with the highest evaluation capacity not only had dedicated BECs but also often had other core M&E staff beyond the BECs. For example, one BEC who described herself as an M&E expert reported that her bureau has a team of four dedicated to M&E, as well as contractors embedded within their office who do five or six evaluations per year.

The more capable bureaus were also more likely to have program officers with M&E capacity, which is largely a product of working in bureaus that do many evaluations and therefore prioritize M&E training and capacity building. In these cases, program officers *“play an important role in developing evaluation SOWs and evaluation questions, helping piece together what evaluations will be done in which years, and what criteria will be used for selecting them. They try very hard to make sure that evaluation is not just seen as a checkbox exercise.”*

Key informants reporting medium evaluation capacity in their bureaus typically had a full-time BEC and program officers trained in certain functions, such as reviewing M&E plans, but no staff with evaluation expertise. A BEC from one of these bureaus said that her bureau does one or two external evaluations per year but does not have the capacity to do more than that at their current staffing level.

Bureaus reporting low evaluation capacity were more likely to have a part-time BEC with limited M&E experience who spends relatively little time on evaluation-related tasks and program officers with little M&E experience. A BEC from one of these bureaus noted that the evaluation capacity in his bureau was *“virtually non-existent. We have to hire contractors for everything, and because there is no in-house evaluation expertise, it is difficult for staff to critically look at the products we get from them.”*

Conclusion 14

The wide disparity in the BEC position in terms roles and responsibilities, training, experience, and capacity constitutes a significant challenge to achieving widespread compliance with the Evaluation Policy and institutionalizing good evaluation (and MfR) practice within the Department. The BECs’ status, moreover, appears to track closely with overall bureau or independent office evaluation capacity being both a reflection of and contributor to bureau or office capacity. The BECs who are the most capable also tend to be full-time BECs who have a wide variety of M&E responsibilities within the bureau or office that include both evaluations and managing for results activities.

The resources provided by F—including tools, courses, and TA—are helpful; however, several BECs are beginning from such a low capacity level that existing F support struggles to compensate for the deficits they face. In such cases, the prospects that the bureau meets all of the DoS Evaluation Policy requirements are relatively slim.

DoS bureaus and independent offices vary dramatically in function, organization, resources, and staffing, and this often translates directly to the resources and staffing devoted to the BEC position and the scope of the BEC’s responsibilities. Evaluation capacity with bureaus and offices tends to rest in individual BECs rather than institutionally within the bureau or office with adverse implications for policy adherence or sustainability.

4.3.5 Capacity and Evaluation Implementation Challenges

As discussed below, key informants reported several main challenges in meeting Evaluation Policy requirements related to bureau evaluation capacity and designing and implementing evaluations.

Staffing and Technical Expertise

Key informants noted a variety of challenges they face conducting evaluations within their bureaus. At the top of the list is the lack of technical acumen in M&E. According to one BEC, for example, *“We don’t have enough of a background in the field of evaluation to think of what kinds of questions to ask, which kinds of projects to evaluate, or what to evaluate within projects.”* Another BEC explained that she was once on rotation for six months, during which time all evaluation activity in her bureau stopped because there was no one else in the bureau that could do the work.

A handful of key informants and focus group participants reported that BEC and staff turnover was a problem, including one key informant who stated that *“some bureaus use foreign service officers for the BEC function who are only here for a couple years before going overseas.”* One BEC interviewee is a foreign service officer who was appointed to the BEC position with no prior evaluation experience after 10 years in the field. Another key informant reported, *“Evaluation capacity can wane very quickly. A bureau might have people that have some ability to manage evaluation contracts, but if they leave and are not replaced, capacity goes down to zero.”*

A burdensome hiring process, made lengthier by a department-wide hiring freeze, has compounded these staffing challenges. One BEC recounted a particularly stark example of this challenge, noting, *“In the fall of 2015 we went through the recruiting and interview process and selected an evaluation expert to hire but getting her through the top-secret clearance process took over a year and by the time it came through in 2017, the candidate was not available due to a family emergency.”*

Funding Constraints

Seven KIIs cited funding as a major challenge, given that program and bureau funding levels and access to funds vary widely across bureaus, while the one evaluation per year requirement is the same for all bureaus. Multiple key informants reported that they do not believe evaluation is a good use of their bureaus’ limited funds. One key informant said, *“As evaluations can often cost as much as \$100,000, spending up to 10 percent of our budget on a single evaluation does not make much sense.”* Other key informants further noted that some bureaus implement several multi-million-dollar projects while other bureaus do not implement any such projects but still must comply with the same Evaluation Policy. One key informant estimated that the programs his bureau implements are typically between \$25,000 and \$300,000, with a range up to about \$500,000, and noted that his bureau does not find it cost effective *“to spend \$100,000 to evaluate a \$300,000 project that occurs over the space of six to eight weeks.”* Another key informant pointed out that, typically, bureaus must take money from funds that have been established for other purposes, because there is typically no earmark for evaluations.

The above quotes should be understood in the context that F has provided guidance on conducting low or no-cost evaluations as part of its suite of tools and resources. However, as noted above, this guidance has not been incorporated into any of the F/BP training courses and largely relies on bureau initiative to seek, access, and adopt the guidance.

Program Relevance/Applicability

Several KII and FGD participants consider the Evaluation Policy to be less relevant to their bureaus’ programs because of the difficulty evaluating them, especially regional bureaus’ diplomatic engagement activities, the effectiveness of which is inherently difficult to measure. Observed one BEC, *“We don’t deploy 50,000 vaccines to Tanzania. What we’re doing is more nuanced and complicated than that and*

the impacts are generational in nature. We are using resources to buy influence and whether there is traction on that influence is a hard evaluation question. Bureau leadership wants to know that we pushed forward the objectives of the bureau, and there's no counterfactual for that." Another key informant questioned how one can effectively evaluate, for example, an African summit or nuclear trainings when, in the latter case, countries do not want their nuclear facilities to be visited.

Bureaucracy

Nine KIIs indicated that complicated bureaucratic contracting requirements are a major hindrance to evaluation implementation, including a cumbersome SOW review process, inefficient indefinite delivery/indefinite quantity (IDIQ) procurement regulations, and an evaluation funds competition that some perceive as overly rigid. (F has stated its intention not to continue with the evaluation funds competition.) Regarding writing SOWs for the funds competition, a BEC remarked that the process is overly nitpicky and iterative for someone who is *"often being pulled in 75 different directions."* Another informant cited managing contractors as a persistent challenge. She explained, *"For example, each contractor tends to subcontract a lot of the work rather than use their own full-time staff, so we have to continually explain how things work within the State Department and within the bureau, which is critical for the evaluation and its findings and recommendations to be put into the appropriate context."*

Conclusion 15

The large diversity of evaluation resources and capacities within the DoS bureaus and independent offices makes devising an Evaluation Policy that is both rigorous and reasonable across all bureaus and independent offices receiving FA funds a challenge. In practice, it also guarantees that some bureaus and offices will struggle to meet its requirements, while others will meet them more easily.

4.3.6 Senior Leadership Buy-In

Eleven (11) KIIs reported a high level of senior leadership buy-in/support for evaluations in their bureaus. In these bureaus, leadership tends to take an active role in advocating for evaluations as an important source of data to strengthen the bureau's knowledge base, improve program learning and implementation, and increase evidence-based decision-making. According to one of the senior leaders interviewed, *"Leadership wants to see what works and how it works. It is important to know what programs are effective, so we can learn from experience and feed evaluation results into programs both in terms of structure and internal management. Evaluations have helped us improved our grant management in this regard."*

In support of these interests, leadership in these bureaus not only prioritizes funding for evaluations but also uses strong evaluation results as an evidence base to lobby for additional program funding. In addition, leadership in these bureaus actively pushes for formalizing and standardizing bureau M&E systems. One key informant reported that in her bureau, *"Leadership has asked us to create standard operating procedures for M&E and has emphasized that M&E needs to be part of the normal processes within the bureau."* Also indicative of this type of prioritization, one BEC reported that her bureau *"has integrated communication with grantee M&E teams into program officer travel protocols. This one small line item is indicative of the importance leadership puts on M&E. The message emphasizes to program officers, 'How should I be involved in doing M&E?'"*

Notwithstanding some bureaus' leadership support for evaluations, nine key informants described leadership buy-in in their bureau as somewhere on the spectrum of low to medium, and they identified a number of factors adversely affecting it. The factors that were raised with relative frequency by key informants are described below.

Timing of Evaluations Is Not Compatible With Decision-Making Timeframes

One challenge in improving senior leadership buy-in for evaluations is that they see evaluations as a time-consuming process in which information on program results or impact is not readily available for consumption, thus discouraging them from promoting evidence-based decision-making. As one FGD participant noted, *“The nature of evaluation is frustrating because it’s not like you do an evaluation, immediately find out that a program is working, and decide then to only implement that program from now on. Instead, evaluation is for building an evidence base incrementally. You might have to re-test an activity in another environment to see if effects are still there, etc.”*

Cost of Evaluations

The expense of evaluations relative to the size of project budgets can be a deterrent to gaining senior leadership buy-in and support for evaluations. Often, leadership believes the benefits of evaluations are outweighed by their cost, and money can be better spent on program activities or other expenses. According to one key informant, *“Generally our evaluation funding tends to only go towards flagship programs. If we want it to go towards anything else, we really have to justify it and push for it, and leadership doesn’t usually find those arguments very compelling. They would rather give \$500,000 directly to program activities that help people rather than directing it towards evaluation.”*

Length of Evaluation Reports

Senior leadership may not want to read lengthy evaluation reports, which hinders their interest in learning and understanding the evaluation process and the benefits of program evaluation. (Notwithstanding this obstacle, the meta-evaluation did not find poor readability, including length and accessibility, of evaluation reports to be a serious issue.)

Evaluations Are One of Many Considerations in Decision-Making

Senior leadership often has to weigh many factors, including political considerations, in making programmatic decisions that often outweigh evaluation results. As a senior bureau official noted, *“The political imperative and ambassadors pushing their agenda from their bilateral engagement often can dictate the strategic planning of issues rather than evidence that shows a certain strategy or program can be more productive than others.”*

Perceived Irrelevance of Evaluations to Bureau Programming

Several KII and FGD participants reported that they consider evaluations irrelevant to initiatives in their bureaus, especially regional bureaus’ diplomatic engagement activities. For example, one FGD participant reported that his bureau works with more than 50 embassies overseas, and that senior leadership does not see evaluation as applicable to this kind of work in the same way it would be to a traditional foreign aid or development assistance program.

Lack of Evaluation Understanding

Senior leadership can lack a full understanding of the purpose for doing evaluations, worrying (as discussed above) that negative evaluation findings will adversely impact their funding rather than seeing evaluation as a tool to improve program learning and implementation. A FGD participant reported that *“Lots of bureau leadership view evaluations as more of a report card and don’t want to post their evaluations because they see evaluation like an audit that could have resource implications. It’s human nature to protect resources.”* Furthermore, a lack of awareness of the programmatic benefits of evaluation deters senior leadership from committing funding and staff to manage and coordinate evaluations. Interviewees also reported a lack of technical evaluation experience and capacity at the senior leadership level, which contributes to misunderstanding on their part as to the purposes and potential benefits of evaluations.

Apathy or Lack of Regular Evaluation Involvement

In many instances, bureau leadership supports evaluation only insofar as it is a checkbox to fulfill the Evaluation Policy's requirements or a tool to justify budget requests, while showing little interest in the programmatic benefits of evaluation or engagement in the evaluation process. The appointment of inexperienced, low capacity, and/or part-time BECs is perhaps one of the principal manifestations of this phenomenon. A key informant summarized perceived senior leadership apathy in her bureau as follows: *"Leadership gives me broad authority to do evaluations, and I use this as leverage with program officers. However, it is not as if leadership actually cares whether they do it; it is not at the forefront of their minds."* Another FGD participant reported that her bureau leaders are *"not holding us accountable to reflect upon recommendations and implement them."*

Evaluations Are Not a Priority

Some bureau leadership sees the importance of doing evaluations and would like to prioritize it, but other priorities and considerations can take precedence. When push comes to shove, evaluations fall in importance and are superseded by operational, political, or other concerns—despite leadership's best intentions. A senior bureau leader came back to this point repeatedly during the KII:

"There is socialization among bureau leadership at the BEC level but at the DAS level and above, there is no socialization...Leadership is the key to institutionalizing the evaluation policy. They don't think it would be effective to do direct outreach to assistant secretaries, as it is very hard to get on their schedules, and typically deputy assistant secretaries are the kind of chief operating officers of the bureaus that make the machinery work. Currently, however, socialization is coming from the BEC up to the DAS and then up higher rather than the reverse. The executive director, DAS, and [principal deputy assistant secretary (PDAS)] are the best levels to target to effect cultural change."

Conclusion 16

Key challenges in meeting the requirements of the Evaluation Policy include a lack of M&E staff and technical expertise among bureau and independent office evaluation staff, funding constraints, bureau programs that do not easily lend themselves to evaluation, bureaucratic obstacles to contracting external entities to do evaluations, and, perhaps most importantly, a lack of leadership buy-in. The socialization of bureau and independent office leadership is a significant and enduring challenge to creating an evaluation culture within the Department.

4.3.7 Change in Evaluation Practice Since 2012

With only two exceptions, key informants reported that the evaluation practice in their bureaus has improved since 2012, and many believe that the Evaluation Policy (initially) and FATAA (subsequently) have been driving forces in this improvement. Specific improvements in evaluation practice cited by key informants include those described below.

Improved Staff Capacity

Staff capacity in evaluation has improved. The creation of the BEC position along with the F evaluation tools, F TA, and the evaluation CoP have all made important contributions to this outcome. One key informant, for example, reported that program staff in her bureau now have a much better understanding of evaluations, why they are important, how they add value, and how they benefit programming. In addition, the technical capacity to perform M&E-related tasks has improved significantly, including many bureaus hiring M&E experts who increased the capacity of program officers in their bureaus.

Increased Number of Evaluations

Overall, bureaus are now doing more evaluations than they did prior to the first Evaluation Policy in 2012. (Certain bureaus had been doing evaluations prior to the Evaluation Policy and have continued at pace since, but they are the exception.) On the high end of the evaluation spectrum, but representing a cross-bureau consensus, one key informant recounted,

“Before 2012, we had only done one evaluation in 10 years, but have done 13 evaluations since 2012, so the Evaluation Policy has been critical in incentivizing us and empowering us to do more evaluations. In the past, we did not get buy-in from program officers because it was not a high priority, because it was not mandated. Without the Evaluation Policy, we would not be anywhere near where we are today.”

Higher Quality and More Useful Evaluations

The consensus among key informants is that the Department is not only doing more evaluations, but also doing higher quality and more useful evaluations, with better structured SOWs, team members with better technical and subject matter qualifications, and more actionable recommendations. According to one key informant, *“Bureaus are becoming more aware of what evaluation consumers want and need. We are seeing far fewer train wrecks in terms of poorly structured SOWs so that the contract can deliver what is in the SOW.”*

Increased Focus on Program Learning

Improved evaluation practice within the Department has coincided with increased focus on using recommendations for program learning and improvement. Multiple key informants reported that their bureaus have created systems for tracking evaluation recommendations. An interviewee reported that after a recent evaluation, *“We went over all the recommendations, ranked them based on level of agreement, articulated how they would treat them, and created an action plan including a timeframe for addressing them.”* Another noted, *“We are trying to institutionalize the idea that bureaus will carry out recommendations or have documented reasons why not and to go into evaluations with this expectation.”*

The use of “recommendation trackers” is a best practice worth emulating across the Department, but there is no evidence that this practice is common across the bureaus at this point. The bureaus using the recommendation tracker tend to be those at the higher end of the capacity spectrum. As discussed above, F has created tools to assist bureaus in implementing and following-up with recommendations as part of its suite of evaluation tools, but it has not integrated these tools into any of its evaluation courses, and it is left largely up to the discretion of the individual bureaus and BECs to seek out and adopt this tool, as well as other F tools.

More Capacity-Building Resources Available

The noted increase in bureau staff M&E capacity comes as F has significantly ramped up its capacity building tools, trainings, and technical assistance—the uses of which are explored in more depth under EQ 4. Nonetheless, it is worth mentioning here that key informants see F as having played an important contributory role in increasing Department evaluation capacity via its courses, tools, and technical assistance.

Increased Buy-In from Senior Leadership

Notwithstanding persistent challenges in securing senior leadership buy-in, several key informants noted increased support for evaluations among bureau leadership since the issuance of the Evaluation Policy and then again after the passage of FATAA. While leadership support still falls far short of where it needs to be, there is now a base of support on which to build that did not previously exist. According to one key informant, *“We have ramped up our evaluation practice in the last few years, and it has helped us to have*

buy-in from leadership. An advantage of the evaluation policy is that we can point to it as a kind of leverage to motivate people to take evaluation seriously. This includes not only leadership within our bureau but outside as well.”

Conclusion 17

The DoS Evaluation Policy has made an important contribution to how bureaus and independent offices approach evaluation and has provided a previously absent impetus within the bureaus and independent offices to take a more proactive approach to evaluation. Evidence for this is found in the numerous changes in evaluation practice within the Department noted by key informants, including improved staff M&E capacity, principally manifested by the institutionalization of the BEC position; an increased number and quality of evaluations; an increased focus on using recommendations for program learning and improvements; the availability of evaluation tools and resources; and (despite the challenges listed above) increased support from senior bureau and office leadership.

4.4 EQ 4: To what degree have bureaus used F’s capacity building tools, classes, papers, and other technical assistance to increase evaluation capacity and improve evaluation?

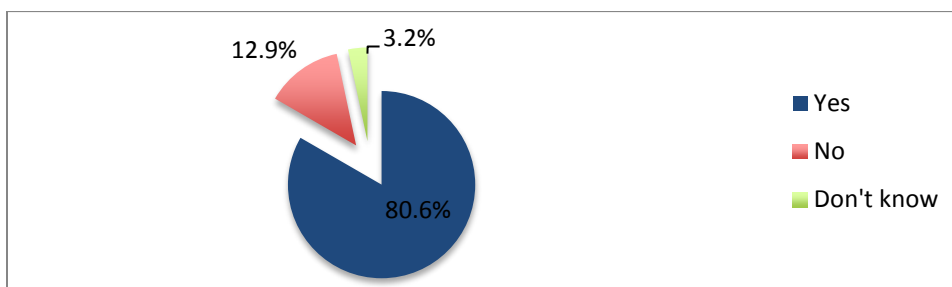
Overall Findings

Most respondents have used one or more of the F evaluation tools, with the evaluation design matrix, SOW toolkit, online resources, and Independent Government Cost Estimate (IGCE) tools receiving the highest favorability ratings among respondents. The demand for the F-designed/BP-funded evaluation courses is high within DoS bureaus and independent offices, and respondents generally believe the courses are effective in building DoS evaluation capacity.

The majority of DoS evaluation staff have consulted directly with F, participated in the evaluation CoP, and used CoP resources, although participation in the CoP is not consistent overall. Most DoS evaluation staff that have consulted with or received TA from F found those consultations useful.

4.4.1 Familiarity with F Evaluation Tools

Figure 9. Familiarity with F Evaluation Tools (N=46)



As seen in Figure 9, over 80 percent of survey respondents are familiar with the F evaluation tools. Thirty-seven (37) or more of the 46 informants (80.4 percent) who responded to this question have used each of seven tools—the SOW toolkit, the dissemination summary template, suggestions for reviewing reports, low/no cost evaluations, the newsletter, the evaluation wiki, and the IGCE—while 29 of the 46 respondents (63.0 percent) have used the evaluation design matrix.

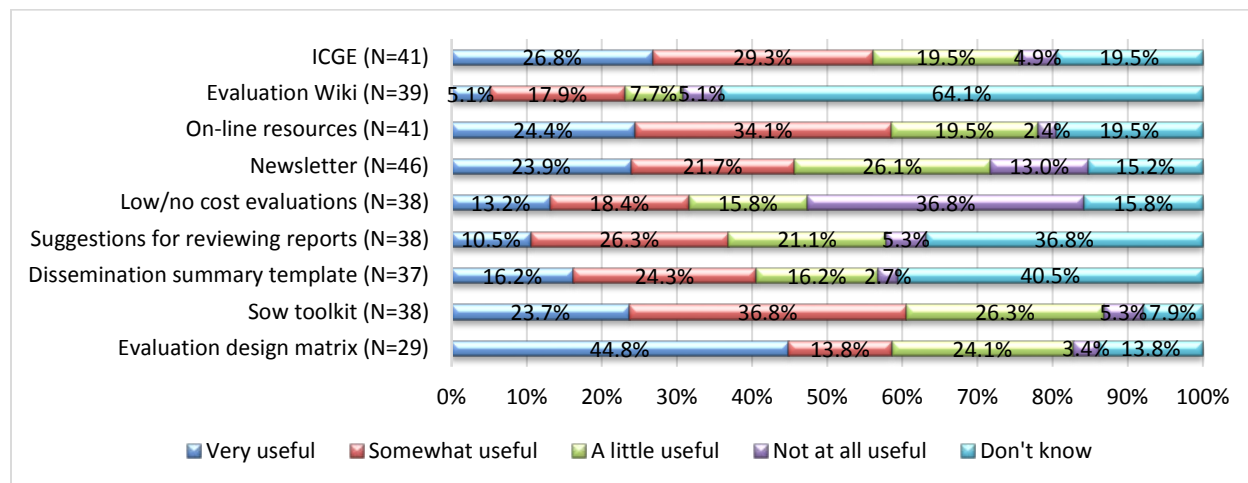
4.4.2 Use and Usefulness of F Evaluation Tools

The evaluation design matrix received the highest approval rating from among the nine tools followed by the SOW toolkit, online resources, and the IGCE tool, each of which was rated as either “very useful” or “somewhat useful” by 50 percent or more of users (Figure 10). Less than one-half of respondents, however, did not find the following tools to be either very or somewhat useful: evaluation wiki,

newsletter, low/no cost evaluations, suggestions for reviewing reports, and dissemination summary template. The tool for conducting low or no-cost evaluations, moreover, was the only one rated as “not at all useful” by a significant number of respondents (36.8 percent).

In the KIIs, 14 key informants reported that the evaluation tools provided by F are useful and are used often or to some extent. Several key informants reported that the F toolkit is very helpful, particularly noting the utility of the SOW template and evaluation design matrix.

Figure 10. F Evaluation Tools and Their Perceived Usefulness⁸



A BEC with significant M&E expertise reported that he has used the evaluation design matrix and the SOW template, and that he finds both of them helpful: *“The biggest weakness for SOWs in the past was that they were asking in a generic way for traditional approaches when perhaps they were not appropriate, and the tool addresses this a bit.”* He added that the SOW template is useful because it is *“written in a way that I can check back to it when writing a SOW to see if I have hit the necessary sections.”* Regarding the evaluation design matrix, he said, *“It is helpful if you have never written an evaluation SOW or designed an evaluation, and it is helpful in getting program officers to think through which questions to ask and which data collection methods to use. Program officers often use it to facilitate meetings.”* He added, however, that the part on writing research or EQs is inadequate: *“If I have to write evaluation questions or if grantees have to write evaluation questions, I often provide resources from other sources because they provide examples that show better in a step-by-step way how it should be done.”*

As a brief aside on the issue of providing examples, the ET interviewed a number of key informants who expressed a desire for more examples demonstrating how a particular tool can be implemented or adapted. For its part, F has requested examples with the intent of sharing them, but bureaus have been reluctant to share, which helps explain the scarcity of practical examples cited by key informants. The bureaus’ reluctance to share reflects a recurring theme during the KIIs that the Department possesses an insular culture that makes sharing or cooperating across bureaus and offices challenging.

Another BEC said that the evaluation design matrix *“is helpful in getting program officers to think through which questions to ask.”* In another KII, respondents reported that they use F tools to help program officers understand the importance of program design and how poor program design makes it difficult to

⁸ Other F tools cited by survey respondents, but which are not included in Figure 31, include sample evaluation reports and SOWs; workshops and conferences; evaluation management schedule; evaluation scope and method chart; developing indicators tools; and trainings. All but two were cited a single time with the exception of sample evaluation reports and SOWs, which were cited twice.

do quality M&E. A full-time BEC, who is the only one doing M&E in his bureau, described himself as “a very avid consumer of F products. I have read and looked at all their resources that are online and used the tools and sample documents.”

In another seven KIs, key informants reported that the F evaluation tools are either not useful or little used, in some cases because the tools are not relevant to their bureau’s activities. As one key informant from a regional bureau noted, “They are not tailored for our needs or fit into the context of what we are doing as a bureau.” One interviewee reported that while she does not use the tools some people take parts of the tools and toolkit and tailor them so that they make more sense. Another BEC stated,

“The logic model and M&E template in the F tools are a little too long for program officers to use, and program officers say that they are written to the wrong audience in that they assume that the program officers within the bureau are coming up with the programs, which is not the case in this bureau, where program officers have a less prescriptive programming style.”

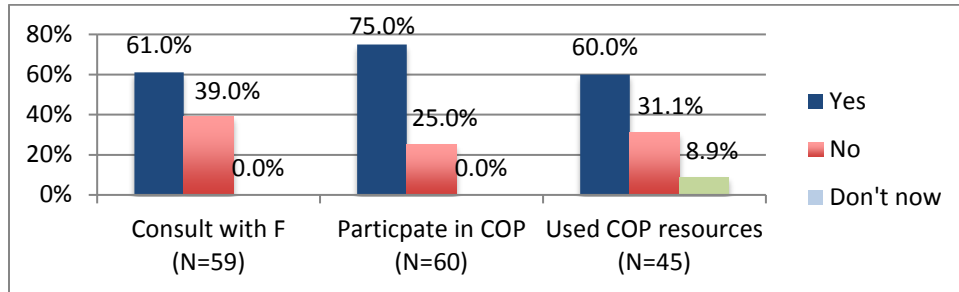
Another BEC tried to use the Program Design and Performance Management Toolkit, which he considered a great idea, but he did not find it useful. He said, “It is incomprehensible to program officers and hard to follow. It is a big document and hard to break down. It needs more examples of actual SOWs, budgets, and what a good report looks like. Otherwise we are just guessing.” Another key informant mentioned that her bureau has its own program M&E guide and tools, including logic models and M&E plan templates, and therefore does not use the F tools. In other cases, some key informants reported that the tools are not tailored to individuals with high technical expertise, so they get their resources from outside sources. Finally, an experienced BEC reported that the F toolkit is insufficient because it provides templates and guidelines without concrete examples.

Conclusion 18
The large majority of Department evaluation staff are familiar with the F evaluation tools and have used one or more of the tools; however, a sizeable number of evaluation staff find the tools to be of limited usefulness. This lack of consensus about the usefulness of F tools reflects the wide variety of experience and capacity found among Department evaluation staff and the difficulty of designing evaluation tools that appeal to an audience with diverse needs. Department evaluation staff, moreover, are largely left to their own initiative to seek out and use or adapt the F tools. The lack of direct training in the tools and the lack of practical examples in how to use/adapt them hinder their use, in the latter case the result of a Department culture that discourages sharing and cooperation across bureaus and offices.

4.4.3 Use of F Consultations/Technical Assistance and Participation in the Evaluation Community of Practice

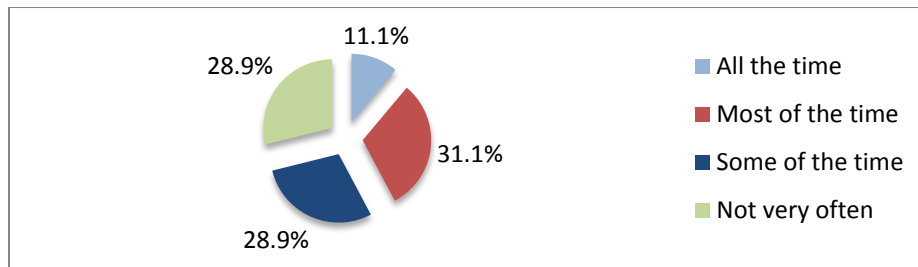
In addition to the F evaluation resources summarized in Figure 10, F offers bureaus additional evaluation services in the form of TA consultations and the evaluation CoP. The latter case holds monthly meetings for BECs and other interested persons and offers a variety of additional evaluation resources for its members. Sixty-one (61) percent of survey respondents have consulted with F at least once (Figure 11), while 75 percent have participated at least once in the CoP. Of the 45 respondents who have participated in the CoP, 27 (60.0 percent) have also availed themselves of additional CoP resources.

Figure 11. Use of Different F Evaluation Resources



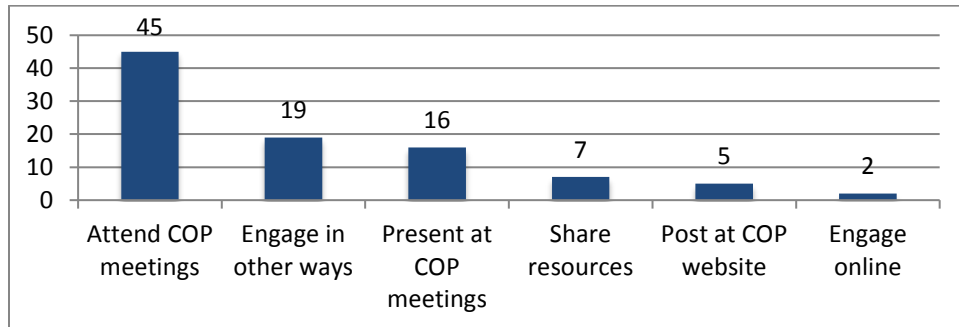
Among those survey respondents participating in the CoP, only slightly more than 10 percent participated “all the time” (Figure 12), with the remaining participants divided more-or-less equally into those who participated “most of the time,” “some of the time,” or “not very often.”

Figure 12. Frequency of Participation in the Evaluation Community of Practice (N=45)



The most common form of CoP participation by far is participating in meetings (Figure 13).

Figure 13. Frequency and Form of Participation in the Evaluation Community of Practice (N=45)



After that, there is a big drop off to the number of CoP members who are engaging in other ways or who present at CoP meetings, with relatively fewer CoP members sharing resources, posting to the CoP website, or engaging other members online.

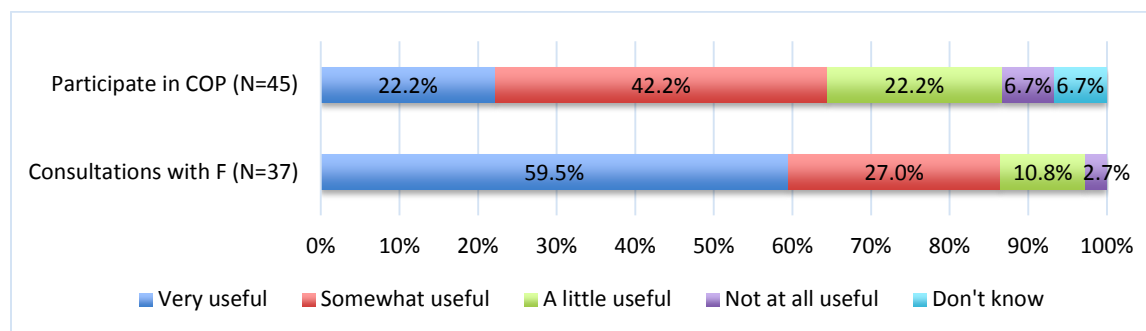
Conclusion 19

The majority of DoS evaluation staff have consulted directly with F, participated in the evaluation CoP, and used CoP resources. Participation in the CoP, however, is more sporadic than consistent and consists largely of attending CoP meetings, with relatively few CoP members engaging with or contributing in other ways to the CoP.

4.4.4 Usefulness of F Consultations/Technical Assistance and the Evaluation Community of Practice

Survey respondents who received consultations from F rate those consultations favorably on the whole, with 59.9 percent saying that the consultations were very useful and another 27 percent saying that they were somewhat useful (Figure 14).

Figure 14. Usefulness of F Consultations and Participation in the Evaluation Community of Practice



In contrast, survey respondents participating in the CoP were much less likely to rate their participation as very useful (22.2 percent), although 42.2 percent did say that their participation in the CoP was somewhat useful. Of further note is that slightly more than one-fifth of CoP participants rated their participation as only a little useful.

Many of the KII participants reported that they have received TA from F and that they found the TA to be very helpful. As one BEC described, *"I have met with F in the past for guidance when doing internal evaluations, which proved to be very useful. Everything I know about evaluation has come from consultations with F and from the BEC meetings or community of practice."* Another key informant noted that F assistance is particularly useful *"in the area of drafting SOWs and doing things such as theories of change, log frames, and so forth."* According to yet another,

"I had a couple of consultations with F and found them very helpful. For example, we avoided doing an impact evaluation because of consultation with F. We didn't know how time consuming it is to conduct impact evaluations and they may not be appropriate for our program. F provided a policy briefing to our program staff and held discussions with us about it."

Only two key informants expressed frustration with F's evaluation assistance, reporting that the SOW process is overly time-consuming and cumbersome. As one interviewee noted:

"We were surprised by how inefficient the process is with F. While the high level of caring from F was good in a way, it was micromanaging. It appeared to us that F cared more about the process than the end result. After we submitted a proposal, F was constantly nitpicking various elements of the SOW. We have to go through all these meetings regardless of how good the SOW is. While we are thankful for the money, the feedback could have been done a lot better. It makes us not want to apply for more funds from F. It was a huge time suck with diminishing returns."

A second key informant said that, while she admires how much F cares about evaluation,

"I sometimes feel like we are being pulled in multiple directions at once. For example, we received an email from F saying the SOW for the evaluation was not sufficient and that the team did not work on it collectively enough, which was not true. We were working in good faith to do a good job; we worked with an evaluation expert and thought between the three of us we were doing well. But now we are discouraged and frustrated."

In the survey, the respondents who found F consultations either only a little useful or not useful gave as the primary reason that the consultations were not ultimately helpful (3 mentions). This was followed by an overly complicated or bureaucratic process for delivering the consultations (2 mentions) and assistance that was not directly relevant to the consultees' needs (2 mentions).

Seven KIIs indicated that the Evaluation CoP is useful, and it is used to a great extent as a resource for knowledge sharing and cross-bureau learning and collaboration. Many key informants said they

contribute to the CoP by presenting some of their work on evaluation and sharing their experiences, a practice that other key informants find beneficial. Another key informant reported that he has attended a number of CoP meetings and ensures that at least one of his M&E staff is at every meeting, as well as that his bureau has posted a number of their resources and templates on the website. He said the most important benefit of the CoP is networking with people in other bureaus, and that he has reached out to other people at the meetings to start dialogue. He added that it is helpful to know who else in the Department is working on evaluations.

BECs asserted that the CoP can be very useful to learn about evaluation, especially for those with less experience. One BEC reported that she *"found the community of practice very helpful because it provides the opportunity to hear and learn new things such as different approaches to program measurement and evaluation."* Another key informant praised the flexibility of the CoP, noting that it posts resources online for those who cannot attend and allows people to call in to meetings. She also said that it is a useful place to get updates, such as the changing IDIQ mechanism or changes in the Evaluation Policy.

Key informants who do not find the CoP useful said either that the information provided there was too basic for experienced evaluators, too advanced for inexperienced evaluators, or was not relevant to their or their bureaus' circumstances. As one key informant noted, *"In the meetings they are so focused on specific examples that I can't see how they are applicable to my office. It is not a good use of my time to spend even one hour per month learning really nuanced things about evaluations for something that is less useful in my work."*

Another key informant reported that the CoP meetings

"...are less useful when they present on more technical issues related to evaluation, as the majority of attendees are at a more basic level. After a couple of meetings, I realized that just because someone is a BEC, that doesn't mean that he or she has experience, and it is often just one of many tasks that form their jobs. The wide diversity makes a community of practice difficult in that it is not easy to meet all the diverse needs of the different participants."

One BEC from a regional bureau noted that he used to go to CoP meetings, but not anymore: *"I wanted to find useful examples that are relevant to the programs we have, but never found these examples."*

Several other key informants reported that the CoP website provides good information and resources, but that it is disorganized and difficult to navigate. Another key informant reported that he usually ends up calling F and asking questions directly, instead.

Online survey respondents indicated the most useful benefit from participating in the CoP is the experience and knowledge participants gain from sharing with their fellow BECs (10 mentions), followed by specific knowledge gained about different evaluation approaches and best practices (8 mentions), increased knowledge about different evaluation requirements and policies (6 mentions), and information about funding sources and opportunities (3 mentions).

Conclusion 20

F's assistance via direct consultants and TA has been effective and offers an important and useful supplement to the standardized nature of the F evaluation tools and F/BP evaluation courses. For the relatively few evaluation staff that did not find F consultations or TA useful, no reasons emerged with sufficient consistency to suggest any underlying problems.

The value-added by CoP participation is limited by two factors. The first is the difficulty involved in structuring CoP meetings to meet the diverse needs of Department evaluation staff. The second is the absence of formal (or structured) networking opportunities it offers to participants.

4.4.5 Usefulness of F and BP Evaluation Training Courses

About 20 key informants indicated that the demand for F-developed and BP-funded evaluation training courses is high within the DoS bureaus, and they are nearly universally perceived as a necessary part of any package of evaluation resources provided to bureau staff. With regards to the courses themselves, key informants in eight KIIs reported that the evaluation training courses are useful and have been used extensively or to some extent. One BEC stated that the design-based training was particularly helpful, as it covered SOW training, and she could make immediate changes to her SOWs after taking the course. She found the other course to have *“less takeaways but is still useful.”* Another key informant reported that she finds the big binder of resources given at F training courses to be very useful and noted that *“when I walk by program officers’ desks, I often see them open and being used.”*

Both respondents in one KII reported that the training courses not only helped them improve their technical capacity but also helped them better understand how to socialize evaluations within their bureau. According to another key informant, *“There is high interest in the evaluation training that is offered, and there is an increase in people wanting to attend the trainings to get up to speed on how to manage an evaluation contractor or work with contractors during an evaluation.”*

Furthermore, multiple key informants found the courses to be useful—if not always entirely well-suited to their needs and level of capacity. Few key informants reported not finding the evaluation courses to be useful. One BEC believes the courses *“are created for beginners and not someone who has done evaluations before. As someone who has spent at least seven years doing evaluations, they are not helpful to me.”* Another key informant reported being very disappointed in the training course he took: *“I was most interested in the performance management portion of it, but the course just reiterated what is in the big binder that F has put out. It included a cookbook of tools. Instead of teaching you how to paint they give you a paint-by-the-numbers template. You can make a nice portrait of a duck with paint by numbers, but you are not going to get a masterpiece.”*

Conclusion 21

The demand for the F and BP evaluation courses is high, reflecting their perceived value among Department evaluation staff. At the same time, their perceived value is limited by an overly standardized approach ill-suited at times to participants’ diverse needs that fails to engender creative application of evaluation practices. The failure to incorporate F’s evaluation tools into the courses is an important oversight that limits their adoption by Department evaluation staff.

4.4.6 Recommendations for Improving F Evaluation Tools and Evaluation CoP

The online survey further asked respondents to recommend additional evaluation tools that F might offer, which are not part of its current suite of tools. The responses covered a range of tools, with the most commonly cited being to provide “more specialized” evaluation tools. Drawing on the KIIs, this is interpreted to mean evaluation tools that are more precisely targeted to specific bureau needs—particularly those with less or more evaluation knowledge and experience. None of the remaining tools were cited by more than two respondents, indicating an absence of consensus as to which additional tools are most demanded by survey respondents.

The survey offered respondents the opportunity to recommend improvements to the evaluation CoP. Two recommendations stand out: 1) providing more skills building or emphasis on best practices/lessons learned/approaches (5 mentions); and 2) opening CoP meetings to broader audiences, including evaluation practitioners and other groups (4 mentions). None of the other remaining recommendations were cited by more than a single respondent—again, indicating a general lack of consensus among survey respondents as to what actions might be taken to improve the utility of CoP participation.

One theme that emerged during the KIIs is that while the CoP is set up ostensibly to facilitate networking among CoP members—and indeed some key informants did mention this aspect of the CoP as one of its benefits—there is no real structure within the CoP to facilitate such networking. Thus, several key informants agreed that networking opportunities among CoP membership were rare and that little actual networking was happening. Members participate in meetings and access CoP resources but do not interact in productive ways, either during CoP events or outside of the CoP. Key informants consistently expressed interest in the idea of transforming the CoP into a more effective networking mechanism. When asked for suggestions of different networking activities, key informants frequently cited something akin to a “happy hour” or another event that brings CoP members together for the express purpose of networking.

At a more general level, several KII participants expressed interest in formal networking opportunities—whether through the CoP or other means—including opportunities to engage directly with more experienced/knowledgeable BECs or other DoS staffers via some sort of on-the-job mentoring program. When asked whether they would be willing to provide such mentoring to their less-experienced colleagues, relatively advanced KII participants were generally open to the idea.

Conclusion 22
There exists demand for F to augment its suite of evaluation tools, but this demand exists largely in general form. Aside from a desire for F to design tools that are better targeted to users’ specific needs, there is no consensus among stakeholders as to what specific forms these augmented tools should take.

4.5 EQ 5: Looking forward to implementation of FATAA, what areas are likely to require more attention or focus from F?

Overall Findings
Additional F support is required to help bureaus implement FATAA, and by extension the revised Evaluation Policy, particularly through promoting buy-in and understanding of bureau leadership, linking evaluation results to decision-making, facilitating the creation of a department-wide M&E support team, providing more and better guidance on how to comply with FATAA, providing guidance to bureaus on how to address common objections or concerns about evaluations, and promoting greater cross-bureau collaboration, including mentoring by more experienced and more capable evaluation staff to less experience and less capable evaluation staff.

4.5.1 Areas Likely to Require More Support or Attention From F

KII and FGD participants cited a number of areas where more support or attention from F is necessary to help bureaus meet the requirements of the DoS Evaluation Policy which, as noted in Section 2.0, represents F’s operationalization of FATAA. Below is a summary of the key themes that emerged.

Promote Buy-In and Understanding of Evaluation from Leadership

Several KII and FGD participants believed that F should do more to promote the benefits of evaluation and opportunities for program learning in order to help increase buy-in and understanding of evaluation from senior leadership, or otherwise engage more directly with senior bureau leadership from the DAS level on up. As one BEC explained, *“Leadership sees evaluation as a check box, but leadership needs a more technical understanding of M&E and how it entails more than just compliance with the Evaluation Policy. It would help us do our jobs better if F could message and target all leadership levels.”* Another key informant noted, *“M&E must infuse all high-level meetings between senior leadership and regional bureaus, to where senior leadership asks the regional bureaus about their data, theory of change, and evidence base, how an output is different from an outcome, etc.”* Or, as an FGD participant argued, *“I don’t think we’ve proven that evaluation is a tool that feeds into our work at the leadership level.”* (Again,

F is aware of and working on this issue, as discussed above.)

Facilitate the Process to Link Evaluation Results to Decision-Making

Multiple key informants stated that F should facilitate the process of linking evaluation results to program decision-making. As one explained:

“There are a number of points in the program decision-making process, such as budgets and strategy, where evaluations can play a major role. The Under Secretary should be turning to F and asking them what they’ve seen from the Africa region in terms of learning from its efforts, for example. F needs to be empowered to represent the M&E accountability effectiveness function. There is a lot of money out there that is wasted, so F needs to be the gatekeeper for data-driven decision-making, including funding-related, and ensure that decisions are validated through evaluation results.”

Serve as or Facilitate the Creation of a Centralized M&E Support Team

FGD participants believe F should play a similar role to what it is currently doing, but in a department-wide capacity. One stated, *“We need someone at the centralized level to say, ‘This makes sense at the department level and at the strategic level,’ and they can develop policy that really works for the Department.”* Another FGD participant asserted, *“We need to have a department-wide team that indicates core competencies that are needed to really do evaluations in that bureau based on what top levels have articulated that the goal is.”* Said another FGD participant, *“If F believes they are going to provide something for us, there needs to be a central evaluation capacity that is expert and provides guidance that is applicable and actionable to the State Department.”*

Hire More Evaluation Experts and Other Relevant Staff

Many KII and FGD participants reported that F should bolster its staff with more in-house M&E expertise than it has—and advocate for certain bureaus to increase their own in-house M&E capacity. *“Increased M&E expertise could be an area of improvement for F,”* reported one key informant, *“For example, experts who know survey design, experimental design, how to build the design, etc.”* An FGD participant from a bureau with a high evaluation capacity similarly noted, *“In our bureau we didn’t get serious about evaluation until we had someone full-time dedicated to it and we had internal expertise being applied to it. That’s when our evaluation practice started to grow rapidly.”* Another FGD participant added, *“It must be someone with the requisite methodological skills. It’s a competency issue that has to be written into position descriptions and taken into account for hiring decisions.”* (Note that F has separately indicated that hiring additional staff to bolster its M&E capacity is not practical under current conditions.)

Provide Better Guidance on Evaluation Policy

Several KIIs and FGDs indicated that F should provide even more guidance to bureaus about the requirements of the Evaluation Policy in order to facilitate compliance, notwithstanding the significant efforts F has made to date to do exactly that. For example, there is lingering confusion among some key informants about how the Evaluation Policy applies to them and their bureaus—particularly for bureaus doing diplomatic engagement or other programs whose results do not easily lend themselves to measurement. As one key informant noted, *“We are trying to create a clear understanding of the policy, but in some areas of the policy there is no clarity. We will work with F and understand the policy better.”*

Another area of confusion expressed by a few key informants is how the requirement to evaluate large programs applies to the programs supported by the bureau that tend to be small, with commensurately small budgets. While the Evaluation Policy clearly defines what constitutes a large program for the purpose of policy compliance—*“Large is defined as meeting or exceeding the median cost of programs, projects, or processes for that bureau or independent office”* (p. 5)—some bureaus still appear to be

struggling to operationalize this definition to their specific context.

Moreover, several key informants questioned whether this requirement is cost-effective for programs which, although qualifying as “large” using the above definition, are nonetheless small in absolute terms, such that the cost of an evaluation would constitute a significant portion of the overall program budget. (Again, F has addressed this issue with its tool “Low Cost, No Cost Evaluations for Activities and Small Projects.” As seen above, over 80 percent of survey respondents have used this tool, although about one-third of them do not find the tool to be particularly useful.)

Key informants also expressed confusion regarding the MfR provisions in 18 FAM 300. According to one key informant, for example, *“There needs to be more structured guidance about what bureaus need to do in terms of the program management. Now they’re leaving it up to the bureaus to decide how to implement it. In my bureau we are still debating what a program is and how to define it. This is an example where more specific guidance from F can help settle a question.”*

Regardless of the effort F has put into publicizing and explaining the Evaluation Policy within the Department, precisely how the policy applies in all cases was still not clear to several key informants who hold evaluation roles within their bureaus. As one FGD participant noted, *“They [F] said, ‘Here’s the policy, now figure out how to take the policy and make it happen.’”*

Provide Guidance on How to Address Common Objections or Concerns About Evaluations

Two key informants expressed doubts that the Evaluation Policy applied to their bureaus because political considerations inevitably trump evidence in making programming decisions. Indeed, issues related to the political calculus of evaluations emerged as a fairly consistent theme across interviews. Key informants pointed to a number of political rationales for doing or not doing evaluations including, principally, the fear of negative findings and the resulting loss of funding support and/or status. Along these lines, several key informants pointed out that, with the existing Department culture, there is a widespread lack of confidence that evaluations will be treated as a “learning experience” for the purpose of improving programs and processes and a strong suspicion—despite all avowals to the contrary—that evaluations will be used instead for punitive purposes.

These doubts point to a more general problem that was a recurring theme in the KIIs and FGDs related to all the previously-discussed reasons bureau staff and leadership may object to doing evaluations, which is also a function of the widespread lack of knowledge about or misunderstanding of evaluations. In such cases, several of the key informants struggled to address and overcome these objections—in some cases due to their own lack of knowledge or experience.

Interestingly, in discussing this issue, none of the key informants mentioned F’s resource “10 Things to Know about Evaluation at State,” which is F’s tool to help evaluation staff address such concerns, suggesting that this tool is not yet seen as the go-to resource for addressing such concerns. In the ET’s judgment, moreover, this tool, while informative, is too general and abstract to the point that a less experienced BEC or other evaluation advocate will still struggle to address specific concerns related to the costs and benefits of doing evaluations. In particular, this includes the suspicion that evaluations may be used for punitive purposes, which is a real and reasonable fear, despite agreement in the abstract that evaluations are not audits and that documenting failures is important for learning purposes.

Promote Greater Cross-Bureau Collaboration

KIIs and FGDs indicated that F should do more to promote cross-bureau learning and collaboration. A key informant reported, *“F could do more to facilitate networking, such as creating some kind of class that requires BEC attendance. They should encourage people in bureaus to swap places to work together on evaluations. One of the most useful things that F did was the evaluation conference, which was a two-day event. At this conference I learned a lot about how other bureaus are doing evaluations.”* Remark

FGD participant, “I would like to see a group where we can brainstorm and bounce ideas off each other. The community of practice isn’t promoting networking.”

Provide Additional Training and Technical Assistance

Several key informants cited a need for more training and technical assistance from F. For example, one reported, “It would be nice if there were some type of induction program, training, or resources for new BECs who have limited evaluation experience. I went and did a lot of my own research and found tools in other places, such as a USAID training course on M&E. It would be nice if I did not have to do this outside research to get these basic-level M&E resources.”

Online Survey Results

When asked what additional support F could provide to bureaus, survey respondents cited three types of additional support more frequently than others: 1) providing additional TA and training, including additional guidance on how to comply with the Evaluation Policy where the applicability of the Evaluation Policy’s requirements are less clear (13 mentions); 2) providing more funding or resource support (10 mentions); and 3) supporting outreach to raise awareness and knowledge about evaluations, while also reducing misconceptions, within the bureaus (9 mentions). Other responses closely related to providing additional TA and training include hiring more evaluation experts and staff within F, disseminating evaluation reports, providing instructions on contracting, and detailing F staff to work directly with bureaus to coordinate M&E. These responses indicate a continued strong demand within the bureaus for additional F TA and training in diverse areas beyond what F has provided to date. Lastly, survey respondents indicated the importance of ongoing engagement efforts at different levels within the bureaus to generate support for doing and using evaluations, such as promoting leadership buy-in or supporting outreach to raise awareness and knowledge.

Conclusion 23

Additional F support is needed in multiple areas to help bureaus and independent offices achieve full compliance with FATAA. Priority areas in which additional F support is required are described in the findings.

5.0. RECOMMENDATIONS

Drawing on the findings and conclusions presented above, the ET offers the following recommendations to aid F in furthering the institutionalization of FA evaluation across the Department. The recommendations are organized by EQ, and for each EQ by rough order of priority. The ET has endeavored to limit recommendations to those that have a reasonable chance of being implemented (if not in the aggregate at least in isolation); nonetheless, all recommendations assume the caveat “to the extent conditions and resources allow.”

5.1 Evaluation Question 1

1. Provide More Detailed Evaluation Report Guidance: F should provide more detailed guidance to BECs and other DoS evaluation staff on quality criteria for evaluation reports. This guidance should be structured to serve as a guide both to preparing and reviewing evaluation reports. While F has provided guidance on the evaluation content and quality (e.g., “Evaluation Report Structure and Content” document included among F’s suite of evaluation tools), this guidance is at a general level and could benefit from increased scope and specificity. An example of such guidance might be something akin to an evaluation report checklist similar in content and structure to the checklist used for the meta-evaluation. The enhanced guidance should also make sure to include, among other things, the following areas found to be of poor or fair quality in the meta-evaluation: 1) including all

key report components in both the executive summary and main report narrative; 2) aligning the report with the SOW; 3) providing actionable recommendations; 4) providing a complete and detailed description of both quantitative and qualitative methods, including sample sizes, sampling methods, and data analysis methods; 5) distinguishing clearly between findings and conclusions; 6) considering project effects on women and different social groups; and 7) including conflict of interest declarations.

2. Link to Examples of Good Quality Reports: F should embed links or references to examples of good quality reports, or reports that demonstrate good quality with regards to certain quality criteria, in all of its guidance documents so that relevant DoS staff can see practical examples of good practice. The ET appreciates the challenges F has motivating individuals and bureaus to share reports or examples of their applied evaluation practice; however, F already has a reasonably substantial database of evaluation reports in its evaluation registry, and among these reports are examples of good applied practice that F can reference.
3. Establish Process for the Review of Draft Evaluation Reports: F should consider establishing a process within the Department for reviewing draft evaluation reports. This process ideally would tap into existing Department resources within its bureaus and independent offices, perhaps as part of the mentoring process proposed in Recommendation #16, so as to ease the burden on F for implementing this recommendation.
4. Work with Partners to Improve Evaluation Reports and Evaluation Capacity: F should work with IDIQ partners to increase their emphasis on and capacity for creating quality evaluation reports emphasizing, among other things, the importance of relevant, specific, and actionable recommendations. This would include, at a minimum, sharing with contractors any existing and to-be-developed report preparation guidance. F should also consider including standard language in contractors' SOWs that they provide training and mentoring of bureau evaluation staff to maximize knowledge transfer.

5.2 Evaluation Question 2

5. Develop a Strategy for Engaging Bureau Leadership: F should develop a strategy for engaging bureau leadership on evaluation issues to increase leadership buy-in for conducting and using evaluations. This strategy might, for example, include some combination of the following: 1) identifying existing opportunities, or the creation of new opportunities, for F and/or bureau evaluation staff to engage directly with leadership; 2) creating a network of change agents within the Department among qualified and/or motivated evaluation staff; and 3) creating a standard set of messages so as to ensure consistent messaging on common questions and how to address common objections to doing evaluations. The purpose of the messaging would be to provide rationales to try to demonstrate that the potential benefits of doing an evaluation exceed the potential risks.

With regards to this last strategy, the ET acknowledges that F has created a tool for this purpose ("10 Things to Know about Evaluation at State"), but in the ET's judgment, this tool does not equip BECs and other evaluation advocates to address some of the most serious disincentives to doing evaluations, such as tradeoffs between costs and benefits, difficult to measure program outcomes, doing evaluations of small and/or short duration projects, and, perhaps most importantly, concerns that poor evaluation results will trigger punitive measures. Communicating these more nuanced and complex messages about evaluation will be particularly difficult for less experienced or less qualified evaluation staff or advocates, and existing tools do not equip them to address these legitimate concerns.

6. Develop a Knowledge Management Strategy: F should develop a strategy for implementing a department-wide knowledge management strategy, including the dissemination, management, and use of evaluation results and lessons learned within the bureaus. Given the overlap between the

knowledge management strategy and the strategy for engaging bureau leadership, the two may be integrated into a single strategy. As a resource for this strategy, F should use the findings of this evaluation to develop a strategy for addressing factors hindering evaluation use. The intent here is for F to develop the strategy but then work with others to implement it as resources allow.

7. Provide Guidance on Creating Learning Products to Enhance Outreach of Key Evaluation Takeaways: F should provide additional guidance to BECs and IDIQ partners on strategies for disseminating evaluation results that increase accessibility and readership among evaluation stakeholders. The existing dissemination resource (“Planning for Evaluation Dissemination”) does not provide guidance on how to write or structure “learning products” designed to increase readership and accessibility. Examples of learning products might include short summaries of key evaluation findings or recommendations, summary charts or graphics, highlights of successful interventions or use of evaluation recommendations, videos, etc. As part of the guidance on creating learning products, F can link to learning products produced within the DoS or externally by other organizations.

5.3 Evaluation Question 3

8. Supplement Existing Guidance on Compliance with the DoS Evaluation Policy: F should supplement existing guidance for complying with the Evaluation Policy via easily digestible summary guidance. The DoS Evaluation Policy and F guidelines for implementing the policy provide substantial guidance to bureaus and independent offices on how to comply with the policy. F has complemented this guidance via an involved process of engaging directly with bureaus and offices in one-on-one and group settings. Despite these efforts, moderate to substantial confusion continues to exist within bureaus and offices about how the policy applies to them and what they need to do to comply with it. The ET acknowledges that it will be extremely difficult for F to eliminate all lingering confusion, owing to, among other things, staff rotation and the disinclination of certain staff to pay attention. Nonetheless, the ET believes that creating a summary one-to-two-page document that specifically and in an easily digestible format summarizes the policy and compliance requirements may help at the margin to increase understanding and facilitate compliance.
9. Provide Additional Guidance on BEC Qualifications: F should provide additional guidance on desired BEC qualifications. Existing guidance on the BEC position found in the Evaluation Policy and F policy guidance is general and vague. There thus exists room for F to provide more detailed guidance on desired BEC qualifications, in addition to guidance on BEC roles and responsibilities. The ET acknowledges that F lacks the authority to require BEC qualifications or BEC roles and responsibilities, but it is already providing guidance related to both, thus the ET is recommending that F augment this guidance with more expansive and detailed language that helps bureaus and independent offices think more critically about the BEC role. For example, F might consider creating something akin to an ideal BEC SOW to serve as a template that bureaus and offices can adapt as appropriate.

Note that the implementation of this recommendation will have potential downstream benefits in terms of training and TA of Department evaluation staff. A primary challenge F faces is devising evaluation resources (e.g., tools, courses, CoP) that appeal to a highly heterogeneous market of users. Currently, many users do not find the resources useful either because they are at a too basic or a too high level or are otherwise not suited to their needs. Indeed, this was one of the most frequent observations among key informants with regards to F’s evaluation resources. Thus, to the extent F can move the Department closer to standardizing the BEC qualifications and roles (within existing constraints), this will facilitate over time the creation of evaluation resources that appeal to a wider group of users.

5.4 Evaluation Question 4

10. Integrate F’s Evaluation Tools into Evaluation Courses: F should integrate its evaluation tools into

future evaluation training courses. F has developed an extensive suite of evaluation tools; however, to date it has not integrated these tools into its evaluation courses offered in conjunction with BP. The training courses offer a prime opportunity to introduce Department evaluation staff to the tools and instruct them in their use. F should, moreover, include a provision in the external contractor SOW that requires them to use the F tools in developing and delivering the training courses.

11. Seek Opportunities to Link Evaluation Staff to F's Evaluation Resources: F should seek out and take advantage of opportunities to link DoS' evaluation staff to F's evaluation resources and examples of good practice. F communicates with evaluation staff via a variety of means, such as emails, a newsletter, fora, workshops, training courses, etc. The communication events provide F with a variety of fora that it can exploit to link the evaluation staff to F's resources and examples. In addition, the existing F tools should, wherever possible, include examples of or links to examples of practical tool application. Again, the ET recognizes the difficulties that F has encountered getting evaluation staff and bureaus to share information; however, F has expressed the intent to continue to pursue this information. Meanwhile, there exists a relatively robust set of evaluations from which to draw examples. In addition, F is not limited to examples from the Department, and where examples from other sources exist, F should feel free to link to those as well.
12. Enhance the Benefits of CoP Membership: F should seek to enhance the benefits of CoP membership through additional activities of interest to groups or members. These activities could be segmented by experience level (e.g., affinity groups) to better tailor activities towards the advanced or more basic evaluation experience levels of the staff.
13. Bring BECs and Other Evaluation Staff Together for Networking and Joint Learning: F should create opportunities for BECs and other DoS evaluation staff to meet for networking and joint learning. Potential opportunities include workshops, brown bags, happy hours, etc. While ostensibly serving a networking role, the evaluation CoP has not played this role in practice and, as it is currently structured, is not likely to play this role in the future. Thus, it is necessary to create opportunities for CoP members to meet, socialize, and network.
14. Develop BEC Induction Resources: F should develop a package of resources (e.g., FAQs, how-to tips, short (e.g., 2-hour) induction course, or links to other resources) that it provides new BECs or inexperienced BECs for the purpose of orienting them and making sure that they understand their jobs and where they can find resources or help when they will inevitably need it.

5.5 Evaluation Question 5

15. Offer Evaluation Courses Targeted to Different Audiences: F should provide options for evaluation courses that are targeted to different audiences with different capacity levels. The ET acknowledges that F is constrained in the number and diversity of evaluation courses it can offer. But, within these existing constraints, it should be possible to create offerings that appeal to Department evaluation staff with different levels of evaluation capacity. These could be, for example, one course offered for less experienced evaluation staff and another course offered to more experienced evaluation staff.
16. Facilitate Peer Mentoring Among Department Evaluation Staff: F should help facilitate peer mentoring opportunities among Department evaluation staff. There exists both demand and potential supply for peer mentoring among BECs and other evaluation staff. The ET believes that, to the extent possible, F should seek to offload some of its burden for providing TA to those within the Department who possess both the capacity and willingness to take on some of this burden via peer mentoring. F's role in this recommendation would be to create the framework for peer mentoring and then act as matchmaker by linking supply to demand.

ANNEXES

Annex I: Evaluation Statement of Work

Examination of Foreign Assistance Evaluation Efforts at the Department of State: A Combined Performance and Meta-Evaluation

ACTUAL TASK ORDER STATEMENT OF WORK

Nature and Purpose of the Evaluation

The Office of U.S. Foreign Assistance Resources (F) is seeking a dual design, mixed methods evaluation of its implementation of the U.S. Department of State's (DoS) evaluation policy. Specifically, F wants to examine the quality and utilization of completed evaluations and the current capacity and technical expertise in bureaus and available from F. Using a meta-evaluation design, F would like to evaluate approximately 125 evaluations completed between 2012 and 2016, looking at appropriateness of design, data collection, analysis, and reporting. The performance design portion of the evaluation is twofold, examining utilization and capacity. F would like to know whether or how bureaus use evaluation results and whether there are obstacles to their use. F also wishes to examine bureau capacity and the efficacy of its provision of technical expertise to assist in implementation of the evaluation policy. Finally, taking into account prior findings and requirements of the 2016 Foreign Aid Transparency and Accountability Act (FATAA), what are the likely strengths and weaknesses of bureaus moving forward and where should F focus its assistance to them? F will use the results of the evaluation to answer questions about evaluation quality, bureau utilization of evaluations, and overall capacity and technical expertise, which will help it target areas where further technical assistance to bureaus could make the most difference in the continued implementation of the evaluation policy. In addition to internal learning, F will use the results of the evaluation to respond to questions from Congress and other interested parties in the foreign assistance community.

Background and Current Status of the Effort

During the past several years, DoS has made significant progress in institutionalizing evaluation. F and the Bureau of Budget and Planning (BP) issued a revised evaluation policy in 2015. F has also issued comprehensive evaluation guidance, and a series of papers and tools to help bureaus and offices in designing, conducting, and utilizing evaluations. F and BP have regularly offered courses on managing evaluations and evaluation design and data collection methods for the Department's employees and contractors. F and BP also provide technical assistance to bureaus in the form of consultations on planning for evaluations, developing Statements of Work (SOWs), providing tools and reference materials and reviewing draft reports, as well as developing conferences, courses, and just-in-time workshops. Most bureaus have evaluation coordinators to plan and manage evaluations. The Department has a vibrant Evaluation Community of Practice (CoP), which meets monthly to discuss relevant issues and share information. It has established websites that bureaus can access to seek additional information. As a result of these developments, increasingly more bureaus have been undertaking evaluations of their projects and programs.

This evaluation will focus on the 26 functional and regional bureaus and offices that use foreign assistance (FA) funds. These bureaus have conducted approximately 125 evaluations since 2012. The policy is intended to promote evaluations that increase learning, promote effectiveness, and provide data for decision-making. F has worked to build capacity in bureaus to conduct evaluation since the first formal policy was established in 2012. Based on two years of implementation, F updated the policy in January 2015. Initially, F intended to examine the implementation of the updated policy after a year. In July 2016, however, President Obama signed FATAA into law. While the Department was already doing much of what the law requires regarding evaluation, it is developing a new comprehensive performance management policy that includes evaluation as part of a continuum of planning, managing, monitoring, evaluation, and learning. As a result, F would like to use the current evaluation to establish where we are

and what we are likely to need moving forward. The new policy will be made available at the start of the engagement, in draft or final form.

Objectives

F will use the information gathered from this evaluation to make decisions about how it will shape its efforts to further institutionalize FA evaluation at DoS. The proposed evaluation is designed to address three objectives:

1. The strengths and weaknesses of evaluations as shown by report content. We need to know about the appropriateness of evaluation designs and data collection methods, problems and challenges evaluation teams face collecting data in the field as detailed in limitations, the impact of such challenges on the validity and reliability of the findings, and the presentation of findings. We also need to know to what extent evaluations answered questions with credible evidence and if the recommendations flowed from the data and findings.
2. The utilization of evaluation findings and recommendations. The information on this subject is scant. We do not know, for example, how many bureaus have been using evaluation findings and recommendations, when they use them, and for what purpose. Nor do we know whether the evaluation was tied to a bureau or regional strategy or whether the results informed a strategy or the design of a new program, project or activity. What obstacles, if any, hamper utilization?
3. Building evaluation expertise and capacity in bureaus. The policy requires bureaus with \$1 million or more in funds to perform one evaluation per year at a minimum. It also requires each bureau to have a Bureau Evaluation Coordinator (BEC). We would like to know how many FA bureaus have met the baseline requirements. For those that have, is there a correlation between evaluation quality and particular practices? How much staff power are bureaus dedicating to evaluation? In addition, we are interested in the degree to which bureaus are familiar with and have used tools and education from F and where there are gaps in information supplied or knowledge. How are bureaus positioned to respond to the new requirements of FATAA?

Evaluation Framework and Rationale

We will be using methodologies of a meta-evaluation for the first objective and those of performance evaluation for the second and third objectives. F considered whether it should commission three distinct evaluations (a meta-evaluation focusing on analytical and methodological rigor, and two other evaluations, one of utilization of evaluation and the other of capacity building) or a single comprehensive evaluation covering all three areas. We opted for a single evaluation for three reasons. First, all of these topics are inter-related. The information, ideas, and insights gained in one area will be utilized in other areas as well. Second, three evaluations will pose problems in data collection. For example, they will require interviewing the same officials and managers more than once, which could affect willingness to participate. Third, it will save time and resources to commission a single evaluation.

Evaluation Questions

As the focus of the evaluation will be on three areas of inquiry, evaluation questions (EQs) are listed separately for each area. The EQs will be adjusted and finalized by the contractor and staff from F following contract award. All terms within each question must be clearly defined by the time the evaluation design document is submitted as a deliverable. Bulleted questions under numbered questions are suggested ways to investigate through interview protocols, focus group questions, or other research instruments. They are not added sub-questions.

Strengths and Weaknesses of Evaluation Reports

The main evaluation questions are as follows:

1. What does the structure and content of evaluation reports tell us about the strengths and weaknesses of methodology, design, and analysis in evaluations? To answer this question, evaluators may consider the following questions:
 - Were the evaluation designs and data collection methods used in FA evaluations appropriate for answering the questions? If not, why not?
 - Were the background and purpose of the evaluation explained?
 - Were the methodology and limitations described?
 - Did the findings and conclusions logically follow from the data evaluators presented in evaluation reports?
 - Did recommendations follow from the findings and conclusions?
 - Were the evaluation recommendations specific and actionable?

Questions for Utilization

Questions for utilization of evaluation findings and recommendations are as follows:

2. To what degree are FA evaluation findings and recommendations being used by bureaus, and what factors contribute to their use or non-use? To answer this question, evaluators may consider the following:
 - At its inception, was the evaluation tied to a Functional Bureau Strategy or Joint Regional Strategy?
 - Do evaluation managers consult key decision-makers and program managers about their information needs before drafting SOWs so that the evaluations will generate the needed information?
 - Did key decision-makers and program managers have plans about how evaluation findings and recommendations would be utilized? Do they use some form of recommendation follow up?
 - Were there enough data on the evaluated programs or projects to draw valid findings and conclusions?
 - How were the findings and recommendations communicated to key decision-makers, program managers, and other stakeholders by evaluation managers and evaluation teams? Did they conform to the norms of effective communication?
 - How did bureaus use the findings and recommendations? For example, to: (a) make improvements in the existing program, (b) develop a follow-up program taking into consideration the recommendations of the evaluation, or (c) inform a subsequent strategy, policy, or budget decision? If not, what impediments kept bureaus from using findings and recommendations?

Capacity Development in Bureaus

The main evaluation questions for capacity development in bureaus are as follows:

3. To what extent have bureaus met the baseline requirements of the policy for planned and completed evaluations and staffing and utilizing the BEC position? To answer this question, evaluators may consider the following questions:
 - Do the FA bureaus have BECs and what level are they within the organization? Are their positions entirely focused on evaluation or do they have other duties? What is the percentage of their time devoted to evaluation activities?
 - Are there other staff members engaged in evaluation activities?
 - How many staff members in the bureau have professional training or experience in designing, managing, and utilizing evaluations?
 - Does bureau evaluation staff find their leadership to be supportive of evaluation efforts? What obstacles do bureau evaluation staff face in designing and managing evaluations?
 - Are there any best practices to be derived from bureaus that have made significant headway in evaluation?
4. To what degree have bureaus used F's capacity building tools, classes, papers, and other technical assistance to increase evaluation capacity and improve evaluation? To answer this question, evaluators may consider the following questions:
 - Have F's capacity development efforts focused on the most pressing needs of bureaus?
 - To what extent do evaluation courses meet bureau needs and expectations?
 - Does the technical assistance provided by F help bureaus in improving their products? Did it also enhance their evaluation capacity? If yes, how?
 - Does bureau staff find papers, guidance documents, and tools prepared by F useful? Do they have suggestions for preparing additional materials?
 - To what extent did bureaus use the CoP (now MfR Evaluation) page to seek information for planning, managing, and utilizing evaluations? Did they find the information relevant and up-to-date?
5. Looking forward to implementation of FATAA, what areas are likely to require more attention or focus from F? To answer this question evaluators may consider the following:
 - What level of awareness do bureau staff have of the requirements of FATAA and the new policy?
 - What actions have bureaus taken to develop program designs that meet the requirements of the new policy?
 - How are bureaus collecting baseline data? Monitoring programs?
 - Are bureaus sufficiently staffed to meet the requirements of FATAA and the new policy?

Evaluation Design and Data Collection Methods

The evaluation will follow a mixed-methods approach. It will entail a document review and analysis to lay the foundation for the evaluation; a meta-evaluation of about 125 completed foreign assistance evaluations to investigate the quality of evaluations from the inception of the policy in 2012; and mixed-methods performance evaluations of the Department's capacity for evaluation and its utilization of evaluation results.

The primary objective of the desk review is to gather data to inform the direction and focus of the evaluation, to finalize a work plan and design for the evaluation, and to develop research instruments for its conduct. During the desk review the evaluation team (ET) will:

- Review documents, web sites, evaluation guidance, and a sample of FA evaluation reports to identify critical issues for evaluation;
- Conduct preliminary interviews with F and select bureaus to gather information for developing interview protocols and other research instruments;
- Develop a checklist for examining the quality of analysis and appropriateness of methodology in completed evaluations;
- Develop research instruments (interview protocols, questionnaires, guides for focus groups, and protocols for group interviews) that will be used for soliciting information, ideas, and recommendations; and
- Based on findings, finalize the questions and evaluation design and work plan.

Based on the output of the desk review, the ET will conduct and complete the meta-evaluation of completed reports and the evaluation of bureau capacity and evaluation use by:

- Undertaking and completing the content analysis of evaluation reports using the checklist prepared during the first phase;
- Analyzing the data generated by the checklist to draw tentative findings, conclusions, and recommendations about the analytical and methodological rigor of completed evaluations; and
- Using qualitative data from the meta evaluation analysis to fold into the design for the performance parts of the evaluation.

For the evaluation of utilization and capacity, building on knowledge gained in the review of evaluation reports the consultant will:

- Undertake a survey of program managers, decision-makers in bureaus, and BECs to gather information about the success or failure of evaluation reports, utilization of evaluation findings and conclusions, and evaluation capacity building; and
- Using the survey results, develop and conduct in-depth key informant interviews (KIIs) with program managers, decision-makers in bureaus, and BECs.

In the analysis and reporting phase of the evaluation, the consultant will:

- Prepare tentative findings, conclusions, and recommendations about the three components of the evaluation;
- Present the preliminary findings for discussion in a focus group or other format;
- Draft the evaluation report following the guidance provided by F;
- Submit the draft to F for comments and suggestions; and
- Finalize the report and present findings to groups.

Annex II: Bibliography

List of Documents Reviewed
GAO Report to Congressional Committees - PEPFAR - Agencies Can Enhance Evaluation Quality, Planning, and Dissemination (May 2012)
State of Evaluation - Evaluation - Improving Programming at the Mid-Point (July 2013)
Meta-Evaluation of Quality and Coverage of USAID Evaluations 2009-2012 (2013 August)
State of Evaluation - Diving In - Writing Your First SOW Under the Department's IDIQ Contract (September 2013)
State of Evaluation - Evaluation in Conflict Environments - Challenge and Response (November 2013)
State of Evaluation - Revisiting the Department's Evaluation Guidance (January 2014)
Guidance for Public Posting of Foreign Assistance Evaluation Reports and Summaries (March 2014)
State of Evaluation - Organizational Change for Better M&E - The IRI's Experience (March 2014)
State of Evaluation at State (May 2014)
State of Evaluation - Not Programs - Evaluating Ongoing Activities and Processes (May 2014)
State of Evaluation - Need to go Outside the IDIQ - Decrypting Market Research (July 2014)
Evaluations of Management Activities and Services (September 2014)
State of Evaluation - Using Evaluation Rubrics - Working with Precision and Uncertainty (September 2014)
Evaluation WIKI (2015)
Department of State Evaluation Policy (January 2015)
State of Evaluation - Promoting Gender Equality in State Department Evaluations (March 2015)
Evaluation Design Matrix (April 2015)
Hints for Using the Design Matrix Tool (April 2015)
Writing Answerable Evaluation Questions (April 2015)
State of Evaluation at State (May 2015)
State of Evaluation - Collaborative Evaluations in the Department - Some Hints and Tools (June 2015)
Post-Evaluation Management Response Template (July 2015)
State of Evaluation - Beneficiary Feedback in Evaluation - Opening a Conversation (September 2015)
State of Evaluation - Evaluation Use - Templates for Following Up (December 2015)
The Unappreciative Inquiry Checklist (December 2015)
Guidance for Planning and Conducting Evaluations at the Department of State (January 2016)
State of Evaluation - Funding for Evaluation - The Best Way to Have It on Hand (March 2016)
10 Things to Know about Evaluation at the Department of State (April 2016)
Foreign Aid Transparency and Accountability Act (FATAA) (June 2016)
FATAA Evaluation Requirements (June 2016)
State of Evaluation - Integrating Strategy and Evaluation (June 2016)
Program Design and Performance Management Toolkit (September 2016)
State of Evaluation - Program Design and Performance Management Toolkit (September 2016)
DoS Organizational Chart (October 2016)
State of Evaluation - Evaluation Use at All Levels (December 2016)
Foreign Aid Transparency and Accountability Act of 2016
CoP Evaluation Host Meeting Standards of Practice (2017)
Data Visualization from CA Office of the Comptroller (2017)
Access to Innovation: Utilizing the Census Bureau Linkage Infrastructure for Evidence-based Research (January 2017)
Agenda, DoS Evaluation CoP Monthly Meeting (January 2017)
CoP Presentation, Palestine Security Sector Reform Program Evaluation (February 2017)
USAID Evaluation per ADS 201 and Evaluation Policy (February 2017)
Draft Posting Copy (February 2017)
State of Evaluation - Evaluability Assessments - Maximizing the Bang for Our Evaluation Buck (March 2017)
The Promise of Evidence-based Policymaking – Report of the Commission on Evidence-based Policymaking (September 2017)
Evaluation Registry by Status (October 2017)

List of Documents Reviewed
18 FAM 300 Guidance – Guidance for the Design, Monitoring, and Evaluation Policy at the Department of State (January 2018)
18 FAM 301 Department of State Program and Project Design, Monitoring, and Evaluation Policy
Beneficiary Feedback in the Department’s Evaluations
Diplomacy in Action – Alphabetical List of Bureaus and Offices
Evaluation Recommendation Tracker
Evaluation Report Structure and Content
Expansion of 18 FAM 300: Program Design & Performance Management
Guidance for Preparing SOWs for Foreign Assistance Evaluation
Guidelines for Evaluation Report Structure, Content and Length
Low Cost, No Cost Evaluations for Activities and Small Project
Planning for Evaluation Dissemination
Role of Evaluation Managers During Implementation of Evaluations
SOW Template
State of Evaluation Survey Charts
State of Evaluation Survey Result
Suggestions for Reviewing Foreign Assistance Evaluation Reports
Why State and USAID have Separate Policies for Program Design and Performance Management

Annex III: Evaluation Matrix

Data Collection Methods	Data Sources	Sampling and Sample Sizes	Analysis Methods	Limitations
EQ 1: What does the structure and content of evaluation reports tell us about the strengths and weaknesses of methodology, design, and analysis in evaluations?				
F provided available documents	Set of completed Department of State Evaluation Reports provided by F	72 completed DoS evaluations	<ul style="list-style-type: none"> • Development of indicators for coding, based on EQs • Development of a coding checklist • Review and scoring of completed evaluation reports • Creation of evaluation report database • Statistical analysis – descriptive analysis 	<ul style="list-style-type: none"> • Selection bias from convenience sample of evaluation reports • Limited information
EQ 2: To what degree are FA evaluation findings and recommendations being used by bureaus, and what factors contribute to their use or non-use?				
Survey KIIs FGDs	DoS Staff	66 DoS staff (survey) 38 DoS staff (KIIs) 13 DoS staff (FGDs)	<ul style="list-style-type: none"> • Quantitative analysis of survey data using SPSS • Qualitative analysis using Atlas.ti 	<ul style="list-style-type: none"> • Limited information bias • Recall bias • Halo bias • Low survey response rate
EQ 3: To what extent have bureaus met the baseline requirements of the policy for planned and completed evaluations and staffing and utilizing the BEC position?				
Survey KIIs FGDs	DoS Staff	66 DoS staff (survey) 38 DoS staff (KIIs) 13 DoS staff (FGDs)	<ul style="list-style-type: none"> • Quantitative analysis of survey data using SPSS • Qualitative analysis using Atlas.ti 	<ul style="list-style-type: none"> • Limited information bias • Recall bias • Halo bias • Low survey response rate
EQ 4: To what degree have bureaus used F's capacity building tools, classes, papers, and other technical assistance to increase evaluation capacity and improve evaluation?				
Survey KIIs FGDs	DoS Staff	66 DoS staff (survey) 38 DoS staff (KIIs) 13 DoS staff (FGDs)	<ul style="list-style-type: none"> • Quantitative analysis of survey data using SPSS • Qualitative analysis using Atlas.ti 	<ul style="list-style-type: none"> • Limited information bias • Recall bias • Halo bias • Low survey response rate
EQ 5: Looking forward to implementation of FATAA, what areas are likely to require more attention or focus from F?				
Survey KIIs FGDs	DoS Staff	66 DoS staff (survey) 38 DoS staff (KIIs) 13 DoS staff (FGDs)	<ul style="list-style-type: none"> • Quantitative analysis of survey data using SPSS • Qualitative analysis using Atlas.ti 	<ul style="list-style-type: none"> • Limited information bias • Recall bias • Halo bias • Low survey response rate

Annex IV: Meta-Evaluation Checklist

Descriptive Information About the Evaluation						
1	Evaluation Number:					
2	Name of Meta Evaluator:					
3	When was the date of the evaluation?					
4	What is the sponsoring bureau/department?					
5	What is the topic of the evaluation?					
6	What type is the evaluation?					
7	Is the evaluation single, multi-country, multi-country regional-specific, or multi-country non-regional-specific?					
8	Countries:					
9	Regions:					
10	Is the evaluation independent or done by DoS?					
11	What is the name of the organization that performed the evaluation?					

Column 1	Indicator	Explanation and Scoring	Indicator Source	Score: 0=No; 1=Yes; NA=Not applicable	Sub-Indicator Response	Explanation
	Executive Summary					
1	Does the report include an Executive Summary? (p. 39)	The report should include an Executive Summary at the beginning of the report and before the main body of the report. Score as “Yes” if the report includes an Executive Summary.	State			
1a	List the length of the Executive Summary in the Sub-Indicator Response column.	While there are no guidelines as to the expected/recommended length of the Executive Summary, this is of potential interest to State. If there is no Executive Summary, score as “NA” in the Sub-Indicator Response column and provide an explanation in the Explanation column. For the purpose of counting pages, count partial pages as full pages. For example, if the ES is 6 & 1/3 pages, count the length as 7 pages.				
2	Does the Executive Summary include a summary of the five primary components of the report: 1) description of project, 2) EQs, 3) findings, 4) conclusions, and 5) recommendations?	At a minimum, the Executive Summary should summarize the primary components/sections of the evaluation report, as described below. Score as “Yes” if the Executive Summary includes a summary/discussion of each of these four primary components and the component is specifically identified as such. If the Executive Summary lists findings and conclusions together but does not distinguish between the two of them, score as “No” and explain this in the “Explanation” column. If there is no Executive Summary, score as “NA” and provide an explanation in the Explanation column.	Recommended			
2a	List the excluded components in the Sub-Indicator Response column.	If there are no excluded components, OR if there is no Executive Summary , score as “NA” in the Sub-Indicator Response column.	Recommended			
	Report Structure					

Column 1	Indicator	Explanation and Scoring	Indicator Source	Score: 0=No; 1=Yes; NA=Not applicable	Sub-Indicator Response	Explanation
3	Is the main body of the report between 25-35 pages (p. 24, 39)?	Bureaus and offices should insist that evaluation reports be empirically grounded, readable, and should capture the interest of the reader. As much as possible, reports should strike a reasonable balance between depth and length. When reports are lengthy, e.g., over 50 pages, readers may be inclined to put them off or only casually read them. On the other hand, if reports are too brief, vital information cannot be presented. The best approach is to have the main body of the report (excluding title page, table of contents, Executive Summary, annexes, etc.) between 25 to 35 pages. Detailed information on analysis, data, or research instruments can be placed in appendices. Score as “Yes” if the report length is between 25-35 pages. For the purpose of counting pages, count partial pages as a full page. For example, if the report is 35 & 1/3 pages, count the length as 36 pages.	State			
3a	List the actual length of the report in the in the Sub-Indicator Response column.					
4	Does the report include the five primary components: 1) description of project, 2) EQs, 3) findings, 4) conclusions, and 5) recommendations? (p. 24)	Reports should include, at a minimum, four primary components. The first component should describe the activity being evaluated and list the EQs. The second should present data and findings in a coherent fashion. The third component should be the conclusions, which represent the evaluators’ judgments based on the findings. The fourth component should be recommendations, which should logically follow from the findings and conclusions. Score as “Yes” if the report includes a summary/discussion of each of these four primary components and the component is specifically identified as such in the report. If the report lists findings and conclusions together but does not distinguish between the two of them, score as “No” and explain this in the “Explanation” column.	State			
4a	List the excluded components in the Sub-Indicator Response column.	If there are no excluded components, score as “NA” in the Sub-Indicator Response column and provide an explanation in the Explanations column.	Recommended			

Column 1	Indicator	Explanation and Scoring	Indicator Source	Score: 0=No; 1=Yes; NA=Not applicable	Sub-Indicator Response	Explanation
5	Is the language in the report “readable?” (p. 39)	As far as possible, words/language in the report should be simple, active, familiar, and culturally and politically sensitive. Department officials are usually not conversant with academic jargon and technical expressions; therefore, if they are used, they should be explained in the text. Score as “Yes” if, in the reader's best judgment, the report is accessible to lay or non-expert readers and adequately explains technical jargon or concepts. If scored as “No,” explain why in the “Explanation” column.	State			
6	Is the report organized around the EQs? (p. 39)	Reports should be organized around the key EQs. For each key EQ, the report should have a separate section or sub-section presenting findings and conclusions. The advantage of this format is that readers are able to find answers to key questions easily. Score as “Yes” if the report is organized around the key EQs with findings and conclusions clearly associated with each EQ. Note that the State guidelines do not indicate that recommendations need to be organized by key EQs. Also, this indicator refers to organizing the report around “key” EQs, not sub-questions. Note: It is not expected that each sub-question should have a separate section or sub-section; they may be integrated into the presentation of the findings and conclusions for the key EQs. If there are no EQs in the report, then still score as a “No” but provide an explanation in the Explanation column.	State			
	Evaluation Objectives and Audience					

Column 1	Indicator	Explanation and Scoring	Indicator Source	Score: 0=No; 1=Yes; NA=Not applicable	Sub-Indicator Response	Explanation
7	Does the report specify the evaluation's objectives? (p. 15)	Specification of objectives helps in sharpening the focus of an evaluation and in formulating EQs. It also makes it easier for evaluators to come up with relevant findings and recommendations. An evaluation's objectives should be stated in such a way that they specify how the generated information will be utilized. For example, it is not enough to state that the objective of the evaluation is to examine the performance of an electoral assistance project or assess its impact. It is also necessary to specify, whenever possible, how the findings will be used to improve the performance of the project or how the lessons of the summative evaluation will feed into planning new projects in that area. Score as "Yes" if the report includes a description of the evaluation's objectives/purpose/goals, etc. and how the evaluation results will be used. If the report describes the objectives/purpose/ goals, etc. but does not describe how the evaluation results will be used, indicate this in the Explanation column. Note that a description of how the evaluation results will be used may be found elsewhere in the report, such as in a discussion of the evaluation audiences (e.g., how each audience will use the results). If this is this case, score as "Yes."	State			
8	Does the report describe the evaluation's audiences (p. 16)?	The report should specify who the audiences for the evaluation are. These include principal audiences and secondary audiences, although the report may or may not designate them as such. Score as "Yes" if a description of the evaluation audiences is included in the report. This description may or may not include a discussion of how each audience will use the evaluation results. Note: the description of the evaluation audiences may be integrated into the description of the evaluation's objectives and vice versa. It is not necessary that the description of the evaluation audiences includes ALL possible audiences, just that it includes some discussion of who the audiences are.	State			
Evaluation Questions						

Column 1	Indicator	Explanation and Scoring	Indicator Source	Score: 0=No; 1=Yes; NA=Not applicable	Sub-Indicator Response	Explanation
9	Does the report include no more than six key EQs? (p. 16)	A single evaluation cannot answer too many questions unless the time and resources are abundant. Therefore, the number of key questions in a report should be limited. When a large number of questions are included, the evaluation tends to lose focus, and the quality of data collection suffers, which undermines the validity and reliability of the findings and recommendations. A good rule of thumb is that an evaluation should contain no more than five or six key questions. Score as “Yes” if the report includes no more than six EQs. If the report does not include any EQs, score as “NA” and provide an explanation in the Explanation column.	State			
9a	List the number of EQs in the Sub-Indicator Response column.	If the number of EQs is not clear, the evaluator should use his/her best judgment. If the report does not include EQs, write in 0 in the Sub-Indicator Responses column. If the report does not include any EQs, score as “NA” and provide an explanation in the Explanation column.	Recommended			
9b	List the number of sub-questions in the Sub-Indicator Response column.	If the number of sub-questions is not clear, the evaluator should use his/her best judgment. If the report does not include sub-questions, write in 0 in the Sub-Indicator response column. If the report does not include any EQs or sub-questions, score as “NA” and provide an explanation in the Explanation column.	Recommended			
10	Are the EQs in the report identical to the EQs in the SOW?	The SOW is essentially the contract from which evaluators must work. It is thus important that the EQs included in the SOW are the same EQs assessed in the evaluation and included in the evaluation report. IF the SOW also includes sub-questions, these are counted as well for this indicator. To be scored as a “Yes,” the EQs and sub-questions in the report need not necessarily be word-for-word the same as in the SOW but they need to be nearly or reasonably word-for-word. If the report provides a summary or paraphrasing of the SOW EQs and sub-questions, score as “No” and provide an explanation in the Explanation column. Score as “NA” if the evaluation report does not include the SOW in an annex, if the SOW does not include EQs, or if the EQs in the SOW are unclear or confusing and provide an explanation in the Explanation column.	Recommended			

Column 1	Indicator	Explanation and Scoring	Indicator Source	Score: 0=No; 1=Yes; NA=Not applicable	Sub-Indicator Response	Explanation
11	Does the report answer all of the EQs? (p. 24)	The purpose of an evaluation report is to provide the evaluators' findings and recommendations on each and every EQ. Accordingly, the answers to all EQs should be provided in the report. This indicator does not apply to sub-questions (even if the SOW includes sub-questions) but to principal EQs only. Score as "Yes" if the report attempts to answer all of the EQs. Note that State guidance does not specify whether answers to all questions necessarily need to be included in the main body of the report versus an annex. It is not relevant for this indicator how well the report answers the questions, just that it attempts to answer them. Some reports are clear about how they are answering the questions, while some may require careful reading to determine whether all questions are being answered. If no EQs are provided in the report, score as "NA" and provide an explanation in the Explanation column. If no EQs are provided in the report BUT are included in the SOW, still score as "NA" and provide the appropriate explanation in the Explanation column.	State			
11a	If the EQs are not identical to the EQs in the SOW, does the report provide a justification/rationale for changing them?	After a SOW is issued, it frequently occurs that DoS decides on a revised set of EQs with the ET, whether to better reflect the situation on the ground, to incorporate new information or for a variety of other reasons. If this is the case, then the report should indicate why the EQs covered by the evaluation, and described in the evaluation report, are different from the questions included in the SOW. Score as "Yes" if you scored "No" in the previous question and the report provides a justification/rationale for the change. Score as "NA" if you scored "Yes" in the previous question. If no EQs are provided in the report, score as "NA" and provide an explanation in the Explanation column.	Recommended			
	Methodology					

Column 1	Indicator	Explanation and Scoring	Indicator Source	Score: 0=No; 1=Yes; NA=Not applicable	Sub-Indicator Response	Explanation
12	Does the report describe the research methodology (p. 24)?	The evaluation report should describe the research methodology that was used for the evaluation. This description may be included either in the main body of the report or in a methodology annex. It is not uncommon, for example, for a 1-3 page methodology description to be included in the main body of the report with a longer and more detailed methods section provided as an annex. Score as "Yes" if the report includes a description of the methodology in either the main body or annex of the report. For this indicator, it is not relevant how detailed the methodological description is just whether one is provided in the report.	State			
13	If the methodology includes interviews (either individual or group) and/or surveys, does the report or methodology annex provide the number of interviews or surveys conducted? (p. 40)	The report should provide information on the number of interviews and surveys that were conducted either in the main body of the report or in an annex. Score as "Yes" if the number of both interviews and surveys are included in the report. State guidelines indicate that this information should be provided in an annex to the report but score as "Yes" regardless of whether it is found in the main body or in the annex. Finally, if the evaluation involved neither interviews or surveys, score as "NA." This may be provided with text, or if the report provides a table listing the interviews and a count of the interviews in the table, then this is counted as "Yes." If, however, a similar table simply lists the interviews but does not provide a count, then score this as "No" and provide an explanation in the Explanation column. If the evaluation involved conducts neither interviews or surveys, OR if the report does not describe the research methodology, score as "NA" and provide an explanation in the Explanation column. Unless the report makes a distinction between the two, for simplicity's sake, evaluators are to assume that the number of interviews and surveys conducted and the number of people interviewed or surveyed is the same.	State			
13a	Indicate in the Sub-Indicator Response column whether the number of interviews (individual and group) was included (0=No, 1=Yes).	If no interviews (either individual or group) were performed, score as "NA" in the Sub-Indicator Response column.				

Column 1	Indicator	Explanation and Scoring	Indicator Source	Score: 0=No; 1=Yes; NA=Not applicable	Sub-Indicator Response	Explanation
13b	Indicate in the Sub-Indicator Response column whether the number of surveys conducted was reported (0=No, 1=Yes).	If no interviews (either individual or group) were performed, OR if the report does not describe the research methodology , score as “NA” in the Sub-Indicator Response column and provide an explanation in the Explanation column. Unless the report makes a distinction between the two, for simplicity's sake, evaluators are to assume that the number of interviews and surveys conducted, and the number of people interviewed or surveyed, are the same.				
14	Does the report or methodology annex describe the sampling methods used to select stakeholders to participate in the evaluation?	In the majority of evaluations, evaluators cannot interview or survey all relevant project stakeholders and must, therefore, use a sampling method to select the stakeholders to participate in interviews, focus groups, surveys, etc. The report should describe the sampling methods used to select the participant stakeholders either in the main body of the report or in an annex for both qualitative and quantitative data collection methods. Common examples include purposive (non-random) sampling methods to select KII participants or probability (random) sampling methods to select survey respondents. Score as “Yes” if there is a description of the sampling methods and this description enables the reader to understand at a reasonable level HOW persons were selected and WHY they were selected to participate in the evaluation. If the report mentions that a “census” of the relevant population was included for interviews, FGDs, or surveys, this counts as describing the sampling methods. Score as “NA” if no sampling of stakeholders was required (this will be rare).	Recommended			
14a	Indicate in the Sub-Indicator Response column whether a description of quantitative sampling methods was included (0=No, 1=Yes).	If no quantitative data collection methods were used, OR if no sampling of stakeholders was required , score as “NA” in the Sub-Indicator Response column and provide an explanation in the Explanation column.				

Column 1	Indicator	Explanation and Scoring	Indicator Source	Score: 0=No; 1=Yes; NA=Not applicable	Sub-Indicator Response	Explanation
14b	Indicate in the Sub-Indicator Response column whether a description of qualitative sampling methods was included (0=No, 1=Yes).	If no qualitative data collection methods were used, OR if no sampling of stakeholders was required , score as “NA” in the Sub-Indicator Response column and provide an explanation in the Explanation column.				
15	Does the report include clear language describing the limitations of the evaluation methodology? (p. 24)	This indicator focuses on whether the report describes the limitations of the data collection methodologies used in relatively clear language. 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; lack of baseline data; selection bias as to sites, interviewees, comparison groups; or seasonal unavailability. 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 be transparent about such methodological limitations. The discussion of methodological imitations may be included in the main body of the report or in an annex. Score as “Yes” if the report includes a discussion of methodological limitations. This indicator does not require a judgment of the completeness or quality of this discussion only whether the discussion of the evaluation's methodological or design limitations are included in the main report or in an annex to the report.	State			

Column 1	Indicator	Explanation and Scoring	Indicator Source	Score: 0=No; 1=Yes; NA=Not applicable	Sub-Indicator Response	Explanation
16	Does the report describe the data analysis methods used?	It is useful for the evaluation report to describe the data analysis methods used to analyze the quantitative and qualitative data collected via the methods described in the report. Information on data analysis methods may be found in the main body of the report or in a methodology or other annex to the report. Score as "Yes" if the report or annex explicitly describes the methods used to analyze both quantitative and qualitative data. The description of the data analysis methods need not necessarily describe a specific/different analysis method for each data collection methodology (e.g., sometimes the same analysis method is used for different data collection methods) but if both qualitative and quantitative data collection methods are used, IT MUST include a description of the data analysis method used for each. Of course, if only quantitative or qualitative methods were used, then the report need only describe the analysis method for that. It is not sufficient that the data analysis methods be inferred from the presentation but there must be an explicit description of them to be scored as "Yes" for this indicator.	Recommended			
16a	Indicate in the Sub-Indicator Response column whether a description of quantitative data analysis was included (0=No, 1=Yes).	If no quantitative data collection methods were used, OR if the report does not include a description of the data analysis methods , score as "NA" in the Sub-Indicator Response column and provide an explanation in the Explanation column.				
16b	Indicate in the Sub-Indicator Response column whether a description of qualitative data analysis was included (0="No," 1=Yes).	If no qualitative data collection methods were used, OR if the report does not include a description of the data analysis methods , score as "NA" in the Sub-Indicator Response column and provide an explanation in the Explanation column.				
	Findings					

Column 1	Indicator	Explanation and Scoring	Indicator Source	Score: 0=No; 1=Yes; NA=Not applicable	Sub-Indicator Response	Explanation
17	In the presentation of findings, does the report draw on data from the range of each of the data collection methods used in the evaluation (p. 24)?	EQs should be answered on the basis of data and evidence. One way to measure this is whether the evaluation findings, as presented in the report, draw on all data collection methods and sources used by the evaluation to answer the EQs. This does not mean that each data collection method/data source must receive equal weight in presenting the findings but rather that the report clearly draws on each of the data collection methods and data sources described in the report to answer the EQs. Score as “Yes” if the findings presented in the report include evidence from each of the data collection methods/data sources described in the report or report annexes. Note that it can be difficult at times to determine how documents read are contributing to the evaluation findings; therefore, this indicator does not include this data source.	State			
18	Does the report explicitly consider and explore alternative explanations? (p. 24)	Although evaluators may be asked to look for alternative causes of documented results or outcomes as part of the EQ, it is possible for evaluators to come across such potential alternative causes unexpectedly. If any such causes are found, it is important that report bring these to the reader's attention. Typically, outcomes/results are influenced by a variety of contributory factors both internal to and external to the project (e.g., other projects, political or economic factors, natural events, etc.). The purpose of this indicator is to indicate whether the report explicitly takes into account other possible contributory factors in presenting the findings and/or conclusions. Score as “Yes” if the report discusses alternative possible factors contributing to documented findings/conclusions.	State			

Column 1	Indicator	Explanation and Scoring	Indicator Source	Score: 0=No; 1=Yes; NA=Not applicable	Sub-Indicator Response	Explanation
19	Does the report explicitly include gender considerations in presenting the findings and/or conclusions? (p. 17)	Projects often affect males and females differently, either due to differential access to project activities and/or differential effects on males and females, for a variety of reasons. It is thus valuable for evaluators to either a) disaggregate findings by sex, or b) explain how the project affects women relative to men. Score as "Yes" if the report either disaggregates findings and/or explains how the project affects the different sexes differently. It is not necessary that all findings/conclusions be disaggregated by gender, only those for which gender considerations are relevant. If, in the reader's judgment, consideration of different gender effects is not relevant (e.g., the evaluation assesses a macro-economic project), score as "NA" and provide a short explanation in the Explanation column.	State			
19a	Indicate in the Sub-Indicator Response column whether the report explicitly includes consideration of different social groups in presenting the findings and/or conclusions (0=No, 1=Yes).	Also, of interest to DoS is whether reports also explicitly address whether the project has had different effects on people belonging to different marginal or vulnerable social groups (e.g., religious, ethnic, tribal, etc.). However, as this is less critical than gender, it is assessed here as a sub-indicator rather than primary indicator. Assess this sub-indicator using the same criteria as for gender but applied here to marginal/disadvantaged social groups. If, in the reader's judgment, consideration of different effects for different social groups is not relevant (e.g., the evaluation assesses a macro-economic project), score as "NA" and provide a short explanation in the Explanation column.				
	Conclusions					
20	Does the report clearly distinguish conclusions from findings?	The report should make a clear distinction between findings and conclusions. Findings are factual statements about an activity, project, program, or policy which are based on the evidence. Good conclusions do not merely restate the findings in other words but are interpretations or judgments based on a synthesis of the findings. Conclusions, moreover, are the answers to the EQs, whereas the findings provide the basis for arriving at those answers. Score as "Yes" if the report makes a clear distinction between findings and conclusions. Note that conclusions may be presented in their own section or sub-section distinct from findings, or they may be integrated with findings. In any case, the distinction between the two should be clear.	Recommended			

Column 1	Indicator	Explanation and Scoring	Indicator Source	Score: 0=No; 1=Yes; NA=Not applicable	Sub-Indicator Response	Explanation
21	Do the conclusions in the report follow from the findings? (p. 40)	Conclusions should follow clearly from the findings without introducing new information not found in the findings. Score as “Yes” if findings are judged OVERALL to follow from the findings. This indicator does not require that all conclusions follow from findings but rather that the large majority of them follow from the findings. Score as “NA” if conclusions are not differentiated from findings and you answered “No” to the previous questions. Also score as “NA” if there are not conclusions in the report and provide an explanation in the Explanation column.	State			
	Recommendations					
22	Are recommendations in the report supported by the findings and conclusions? (p. 24, 40)	Recommendations should follow logically from the findings and conclusions found in the evaluation report. If there, can the reader follow a transparent path from findings and conclusions to recommendations? Stakeholders are more likely to adopt evaluation recommendations when those recommendations are based on credible empirical evidence and an analysis that transparently demonstrates why a specific recommendation is a reasonable course of action. Score as “Yes” if the recommendations are judged OVERALL to be supported by the findings and conclusions. This indicator does not require that all recommendations are supported by the findings and conclusions but rather that the large majority of them are supported by the findings and conclusions. Score as “NA” if no recommendations are included in the report or if no findings/conclusions are provided in the report and provide an explanation in the Explanation column. This indicator does not require that BOTH findings and conclusions are included and distinguished in the report; if only one of them is, this is sufficient to score this indicator.	State			
22a	Indicate in the Sub-Indicator Response column the number of recommendations included in the report.	If no recommendations are included, put “NA” in the appropriate column.				

Column 1	Indicator	Explanation and Scoring	Indicator Source	Score: 0=No; 1=Yes; NA=Not applicable	Sub-Indicator Response	Explanation
23	Are recommendations in the report specific? (p. 24)	Recommendations should not only mention what needs to be done, but how it should be done using specific language and avoiding overly general language. For example, the recommendation “improve project management” is overly general and much less specific than “consolidate administrative functions under a single administrative office.” Recommendations that are directed to multiple stakeholders should, moreover, specify which stakeholder should implement which recommendation. Score as “Yes” if, in the reader's best judgment, the recommendations are specific according to this definition. This indicator does not require that all recommendations in the report are specific but rather that the large majority of them are specific. Score as “NA” if no recommendations are included in the report and provide an explanation in the Explanation column.	State			
24	Are recommendations in the report actionable? (p. 24, 25)	Recommendations are “actionable” to the extent that they can be implemented by the USG and its partners in existing circumstances and that the concerned decision-makers have the authority and resources to implement them. Score as “Yes” if, in the reader's best judgment, the recommendations are actionable according to this definition. This indicator does not require that all recommendations in the report are actionable but rather that the large majority of them are actionable. Score as “NA” if no recommendations are included in the report and provide an explanation in the Explanation column.				
	Annexes					
25	Is the evaluation SOW included as an annex to the report? (p. 40)	The report should include a copy of the evaluation SOW. Score as “Yes” if a copy of the SOW is included as an annex to the report.	State			
26	Does the report include either a bibliography or footnotes of documents read for the evaluation? (p. 40)	The report should include either a) a bibliography of documents read as part of the evaluation, or b) footnotes of documents read for the evaluation. Score as “Yes” if a bibliography or footnotes are included in the report. The format or “completeness” of the bibliography or footnotes is not relevant in scoring this indicator. Indicate in the Explanation column whether the report includes a bibliography or footnotes or both.	State			

Column 1	Indicator	Explanation and Scoring	Indicator Source	Score: 0=No; 1=Yes; NA=Not applicable	Sub-Indicator Response	Explanation
27	Are copies of the research instruments included as an annex to the report? (p. 40)	The report should include copies of the research instruments used for the evaluation in a separate annex or annexes. Research instruments include, for example, interview/discussion guides, surveys, assessment tools, etc. Score as “Yes” if the annex includes copies of the research instruments. This indicator does not necessarily require that copies of ALL research instruments are included in the annex, but rather that there is an annex that includes the research instruments. If some instruments are not included in the annex, score as “Yes” but report this in the Explanation column.	State			
28	Is a list of persons interviewed included as an annex or in the main body of the report? (p. 40)	The report should list the names of people who participated in individual interviews either in a separate annex or in the main body of the report. Score as “Yes” if the list of persons interviewed is provided. Note that this indicator refers only to individual interviews or individual interviews with more than one participant. It does not refer to group interviews organized specifically as group discussions (e.g., FGDs). DoS guidelines indicate that this information should be provided in an annex to the report but score as “Yes” regardless of whether it is found in the main body or in the annex. Note that due to sensitivities, in some cases the evaluation may not list the names of the persons interviewed. If the report indicates that this is the case, note this in the Explanation column. If the evaluation did not include interviews (e.g., the report is based on a quantitative survey and not interviews, so there is no requirement to list the names of survey respondents, particularly large surveys), score as “NA” and provide an explanation in the Explanation column.	State			

Column 1	Indicator	Explanation and Scoring	Indicator Source	Score: 0=No; 1=Yes; NA=Not applicable	Sub-Indicator Response	Explanation
29	Is a conflict of interest declaration included in the annex or main body of the report?	It is important that ET 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. It is thus expected that an evaluation will include a declaration as to whether ET members have any conflicts of interest. This declaration is typically found in the annex as a formal conflict of interest statement/form, but it may also be included as a simple declaration elsewhere in the report. Score as “Yes” if a conflict of interest declaration exists in the report. It is not necessary that the declaration is provided for every member of the ET; if it exists for one or more members, that is sufficient to be scored as “Yes.”	Recommended			

Annex V: Online Survey Questionnaire

Demographic Information

1. What Bureau/Independent Office are you in?

- | | |
|---|---|
| <input type="radio"/> Africa Affairs (AF) | <input type="radio"/> Intelligence and Research (INR) |
| <input type="radio"/> Arms Control, Verification and Compliance (AVC) | <input type="radio"/> International Information Programs (IIP) |
| <input type="radio"/> Conflict and Stabilization Operations (CSO) | <input type="radio"/> International Narcotics and Law Enforcement Affairs (INL) |
| <input type="radio"/> Counter Terrorism (CT) | <input type="radio"/> International Organization Affairs (IO) |
| <input type="radio"/> Democracy, Human Rights, and Labor (DRL) | <input type="radio"/> International Security and Nonproliferation (ISN) |
| <input type="radio"/> East Asian and Pacific Affairs (EAP) | <input type="radio"/> Near Eastern Affairs (NEA) |
| <input type="radio"/> Economic and Business Affairs (EB) | <input type="radio"/> Oceans and International Environmental and Scientific Affairs (OES) |
| <input type="radio"/> Energy Resources (ENR) | <input type="radio"/> Political-Military Affairs (PM) |
| <input type="radio"/> Europe and Eurasian Affairs (EUR) | <input type="radio"/> Population, refugees, and Migration (PRM) |
| <input type="radio"/> Foreign Assistance (F) | <input type="radio"/> Science and Technology Adviser (STAS) |
| <input type="radio"/> Global AIDS Coordinator (S/GAC) | <input type="radio"/> South and Central Asian Affairs (SCA) |
| <input type="radio"/> Global Partnerships (S/GPI) | <input type="radio"/> Trafficking in Persons (J/TIP) |
| <input type="radio"/> Global Women's Issues (S/GWI) | <input type="radio"/> Western Hemisphere Affairs (WHA) |
| <input type="radio"/> Other (please describe) | |

2. Are you a Foreign Service, Civil Service or other type of staff member?

- ☐ Foreign Service
- ☐ Civil Service
- ☐ Contractor
- ☐ Other

3. How many years have you worked at in the State Department?

- ☐ 0-5
- ☐ 6-10
- ☐ More than 10
- ☐ Don't know

4. How many years have you worked in your current position?

- ☐ 0-5
- ☐ 6-10
- ☐ More than 10
- ☐ Don't know

5. Are you a Bureau Evaluation Coordinator?

- ☐ Yes
- ☐ No
- ☐ Don't know

6. Does your position (job) description include evaluation as part of the responsibilities of your position?

- ☐ Yes
- ☐ No
- ☐ Don't know

7. Have you previously received formal evaluation training? For example, training in evaluation design, methodology or implementation?

- ☐ Yes
- ☐ No
- ☐ Don't know

8. How many years of evaluation experience do you have? If you have no evaluation experience, please mark "Not applicable."

☐ Less than 1

☐ 1-2

☐ 3-5

☐ 6-10

☐ More than 10

☐ Don't know

☐ Not applicable

On-Line Survey.

Importance of Evaluations and Factors Contributing to their Use

9. How important do you think it is to evaluate your Bureau's/Office's programs, projects or activities?

- | | |
|--|--|
| <input type="radio"/> Very important | <input type="radio"/> Not at all important |
| <input type="radio"/> Somewhat important | <input type="radio"/> Don't know |
| <input type="radio"/> Not very important | |

10. How familiar are you with the evaluations that your Bureau/Independent Office has conducted of its programs, projects or activities?

- | | |
|---|---|
| <input type="radio"/> Very familiar | <input type="radio"/> Not at all familiar |
| <input type="radio"/> Somewhat familiar | <input type="radio"/> Don't know |
| <input type="radio"/> Not very familiar | |

On-Line Survey.

Importance of Evaluations and Factors Contributing to their Use

To what extent do you believe that your Bureau/Independent Office has used evaluation findings, conclusions and recommendations for the following purposes? Please select "ok" to proceed to the questions.

11. Inform your Bureau's/Independent Office's strategy

- | | |
|------------------------------------|----------------------------------|
| <input type="radio"/> Great extent | <input type="radio"/> No extent |
| <input type="radio"/> Good extent | <input type="radio"/> Don't know |
| <input type="radio"/> Some extent | |

12. Inform the design of a new program, project or activity

- | | |
|------------------------------------|----------------------------------|
| <input type="radio"/> Great extent | <input type="radio"/> No extent |
| <input type="radio"/> Good extent | <input type="radio"/> Don't know |
| <input type="radio"/> Some extent | |

13. Improve program, project or activity implementation

- | | |
|------------------------------------|----------------------------------|
| <input type="radio"/> Great extent | <input type="radio"/> No extent |
| <input type="radio"/> Good extent | <input type="radio"/> Don't know |
| <input type="radio"/> Some extent | |

14. Contribute to your Bureau's/Independent Office's learning

- | | |
|------------------------------------|----------------------------------|
| <input type="radio"/> Great extent | <input type="radio"/> No extent |
| <input type="radio"/> Good extent | <input type="radio"/> Don't know |
| <input type="radio"/> Some extent | |

15. Support change management in your Bureau/Independent Office

- | | |
|------------------------------------|----------------------------------|
| <input type="radio"/> Great extent | <input type="radio"/> No extent |
| <input type="radio"/> Good extent | <input type="radio"/> Don't know |
| <input type="radio"/> Some extent | |

16. Question basic assumptions about your Bureau's/Independent Office's policies or practices

- | | |
|------------------------------------|----------------------------------|
| <input type="radio"/> Great extent | <input type="radio"/> No extent |
| <input type="radio"/> Good extent | <input type="radio"/> Don't know |
| <input type="radio"/> Some extent | |

17. If your Bureau/Independent Office has used evaluations for any purpose other than those listed above, please write it in here. Provide up to two additional answers. If your Bureau/Independent Office has not used evaluations for any other purpose, please leave blank and proceed to the next question.

Response 1

Response 2

On-Line Survey.

Importance of Evaluations and Factors Contributing to their Use

18. Overall, to what extent do you believe that your Bureau/Independent Office has used evaluation findings, conclusions and recommendations?

- | | |
|------------------------------------|----------------------------------|
| <input type="radio"/> Great extent | <input type="radio"/> No extent |
| <input type="radio"/> Good extent | <input type="radio"/> Don't know |
| <input type="radio"/> Some extent | |

To what extent do you think that the following have hindered the use of evaluation findings, conclusions and recommendations by your Bureau/Independent Office? Please select "ok" to proceed to the questions.

19. Failure to link the evaluation to Bureau/Independent Office or regional strategy

- | | |
|------------------------------------|----------------------------------|
| <input type="radio"/> Great extent | <input type="radio"/> No extent |
| <input type="radio"/> Good extent | <input type="radio"/> Don't know |
| <input type="radio"/> Some extent | |

20. Lack of clear management support for the evaluation

- | | |
|------------------------------------|----------------------------------|
| <input type="radio"/> Great extent | <input type="radio"/> No extent |
| <input type="radio"/> Good extent | <input type="radio"/> Don't know |
| <input type="radio"/> Some extent | |

21. Low awareness of the evaluation among Bureau/Independent Office staff

- | | |
|------------------------------------|----------------------------------|
| <input type="radio"/> Great extent | <input type="radio"/> No extent |
| <input type="radio"/> Good extent | <input type="radio"/> Don't know |
| <input type="radio"/> Some extent | |

22. Low interest in the evaluation among Bureau/Independent Office staff

- | | |
|------------------------------------|----------------------------------|
| <input type="radio"/> Great extent | <input type="radio"/> No extent |
| <input type="radio"/> Good extent | <input type="radio"/> Don't know |
| <input type="radio"/> Some extent | |

23. Failure to involve stakeholders in the evaluation process

- | | |
|------------------------------------|----------------------------------|
| <input type="radio"/> Great extent | <input type="radio"/> No extent |
| <input type="radio"/> Good extent | <input type="radio"/> Don't know |
| <input type="radio"/> Some extent | |

24. Evaluation design or methodology that is overly complex or rigorous

- | | |
|------------------------------------|----------------------------------|
| <input type="radio"/> Great extent | <input type="radio"/> No extent |
| <input type="radio"/> Good extent | <input type="radio"/> Don't know |
| <input type="radio"/> Some extent | |

25. Evaluation design or methodology that is insufficiently complex or rigorous

- | | |
|------------------------------------|----------------------------------|
| <input type="radio"/> Great extent | <input type="radio"/> No extent |
| <input type="radio"/> Good extent | <input type="radio"/> Don't know |
| <input type="radio"/> Some extent | |

26. Evaluation reports that are too long or otherwise hard to read

- | | |
|------------------------------------|----------------------------------|
| <input type="radio"/> Great extent | <input type="radio"/> No extent |
| <input type="radio"/> Good extent | <input type="radio"/> Don't know |
| <input type="radio"/> Some extent | |

27. Evaluation reports with unclear findings and conclusions

- | | |
|------------------------------------|----------------------------------|
| <input type="radio"/> Great extent | <input type="radio"/> No extent |
| <input type="radio"/> Good extent | <input type="radio"/> Don't know |
| <input type="radio"/> Some extent | |

28. Evaluation reports with recommendations that are not useful or actionable

- | | |
|------------------------------------|----------------------------------|
| <input type="radio"/> Great extent | <input type="radio"/> No extent |
| <input type="radio"/> Good extent | <input type="radio"/> Don't know |
| <input type="radio"/> Some extent | |

29. Inadequate dissemination of evaluation findings, conclusions or recommendations

- | | |
|------------------------------------|----------------------------------|
| <input type="radio"/> Great extent | <input type="radio"/> No extent |
| <input type="radio"/> Good extent | <input type="radio"/> Don't know |
| <input type="radio"/> Some extent | |

30. Evaluation results that come too late to inform decisions

- | | |
|------------------------------------|----------------------------------|
| <input type="radio"/> Great extent | <input type="radio"/> No extent |
| <input type="radio"/> Good extent | <input type="radio"/> Don't know |
| <input type="radio"/> Some extent | |

On-Line Survey.

Importance of Evaluations and Factors Contributing to their Use

31. Overall, what do you think are the most important factors that facilitate the use of evaluation findings, conclusions and recommendations? Provide up to three responses. If you do not have an answer, please leave blank and proceed to the next question

Response 1

Response 2

Response 3

32. Overall, what do you think are the most important factors that hinder the use of evaluation findings, conclusions and recommendations? Provide up to three responses. If you do not have an answer, please leave blank and proceed to the next question

Response 1

Response 2

Response 3

Familiarity with and Adherence to Department Evaluation Policy

33. How familiar are you with the State Department Evaluation Policy?

- | | |
|---|---|
| <input type="radio"/> Very familiar | <input type="radio"/> Not familiar at all |
| <input type="radio"/> Somewhat familiar | <input type="radio"/> Don't know |
| <input type="radio"/> Not very familiar | |

34. In your opinion, to what extent does the management of your Bureau/Independent Office support implementation of the Department Evaluation Policy?

- | | |
|------------------------------------|----------------------------------|
| <input type="radio"/> A great deal | <input type="radio"/> Not at all |
| <input type="radio"/> A good deal | <input type="radio"/> Don't know |
| <input type="radio"/> Some | |

35. To what extent do you think that the Department Evaluation Policy applies to your Bureau/Independent Office?

- | | |
|----------------------------------|----------------------------------|
| <input type="radio"/> Completely | <input type="radio"/> Not at all |
| <input type="radio"/> Partially | <input type="radio"/> Don't know |
| <input type="radio"/> Some | |

36. Overall, to what extent do you think your Bureau/Independent Office has met the requirements of the Department Evaluation Policy?

- | | |
|----------------------------------|----------------------------------|
| <input type="radio"/> Completely | <input type="radio"/> Not at all |
| <input type="radio"/> Partially | <input type="radio"/> Don't know |
| <input type="radio"/> Some | |

37. What would be your priorities for support from F for your Bureau/Independent Office to meet the requirements of the Department Evaluation Policy? Provide up to three responses. If you do not have an answer, please leave blank and proceed to the next question.

Response 1	<input type="text"/>
Response 2	<input type="text"/>
Response 3	<input type="text"/>

On-Line Survey.

Familiarity with and Adherence to Department Evaluation Policy

38. Does your Bureau/Independent Office have a Bureau Evaluation Plan?

- ☐ Yes
- ☐ No
- ☐ Don't know

Bureau Evaluation Coordinator

39. Does your Bureau/Independent Office have a Bureau Evaluation Coordinator?

- ☐ Yes
- ☐ No
- ☐ Don't know

40. How often have you consulted with your Bureau Evaluation Coordinator on evaluation issues? If you are a Bureau Evaluation Coordinator, please mark "Not applicable".

- | | |
|--------------------------------|--------------------------------------|
| <input type="radio"/> Often | <input type="radio"/> Not at all |
| <input type="radio"/> Some | <input type="radio"/> Don't know |
| <input type="radio"/> A little | <input type="radio"/> Not applicable |

41. Overall, how useful have your consultations with your Bureau Evaluation Coordinator been? If you are a Bureau Evaluation Coordinator, please mark "Not applicable".

- | | |
|---------------------------------------|---|
| <input type="radio"/> Very useful | <input type="radio"/> Not at all useful |
| <input type="radio"/> Somewhat useful | <input type="radio"/> Don't know |
| <input type="radio"/> A little useful | <input type="radio"/> Not applicable |

Please indicate which of the evaluation tools and templates you have used and how useful you found them. If you have not used the tool please select "Not applicable". Please select "ok" to proceed.

43. Writing Researchable Evaluation Questions/Evaluation Design matrix template

- | | |
|---------------------------------------|---|
| <input type="radio"/> Very useful | <input type="radio"/> Not at all useful |
| <input type="radio"/> Somewhat useful | <input type="radio"/> Don't know |
| <input type="radio"/> A little useful | <input type="radio"/> Not applicable |

44. SOW toolkit

- | | |
|---------------------------------------|---|
| <input type="radio"/> Very useful | <input type="radio"/> Not at all useful |
| <input type="radio"/> Somewhat useful | <input type="radio"/> Don't know |
| <input type="radio"/> A little useful | <input type="radio"/> Not applicable |

45. Dissemination summary template

- | | |
|---------------------------------------|---|
| <input type="radio"/> Very useful | <input type="radio"/> Not at all useful |
| <input type="radio"/> Somewhat useful | <input type="radio"/> Don't know |
| <input type="radio"/> A little useful | <input type="radio"/> Not applicable |

46. Suggestions for Reviewing Foreign Assistance Evaluation Reports

- | | |
|---------------------------------------|---|
| <input type="radio"/> Very useful | <input type="radio"/> Not at all useful |
| <input type="radio"/> Somewhat useful | <input type="radio"/> Don't know |
| <input type="radio"/> A little useful | <input type="radio"/> Not applicable |

47. Low Cost or No Cost Evaluations for Activities and Small Projects

- | | |
|---------------------------------------|---|
| <input type="radio"/> Very useful | <input type="radio"/> Not at all useful |
| <input type="radio"/> Somewhat useful | <input type="radio"/> Don't know |
| <input type="radio"/> A little useful | <input type="radio"/> Not applicable |

48. State of Evaluation Newsletters

- | | |
|---------------------------------------|---|
| <input type="radio"/> Very useful | <input type="radio"/> Not at all useful |
| <input type="radio"/> Somewhat useful | <input type="radio"/> Don't know |
| <input type="radio"/> A little useful | <input type="radio"/> Not applicable |

49. On-line F DOS resources

- | | |
|---------------------------------------|---|
| <input type="radio"/> Very useful | <input type="radio"/> Not at all useful |
| <input type="radio"/> Somewhat useful | <input type="radio"/> Don't know |
| <input type="radio"/> A little useful | <input type="radio"/> Not applicable |

50. Evaluation WIKI

- | | |
|---------------------------------------|---|
| <input type="radio"/> Very useful | <input type="radio"/> Not at all useful |
| <input type="radio"/> Somewhat useful | <input type="radio"/> Don't know |
| <input type="radio"/> A little useful | <input type="radio"/> Not applicable |

51. Independent Government Cost Estimate template (IGCE)

- | | |
|---------------------------------------|---|
| <input type="radio"/> Very useful | <input type="radio"/> Not at all useful |
| <input type="radio"/> Somewhat useful | <input type="radio"/> Don't know |
| <input type="radio"/> A little useful | <input type="radio"/> Not applicable |

52. If you have used other evaluation tools and templates developed by F or found at the evaluation community of practice not listed above, please write them in here. Provide up to two additional answers. If you have not used any other tools or templates developed by F, please leave blank and proceed to the next question.

Response 1

Response 2

On-Line Survey.

Use of F Tools and Assistance

53. What additional evaluation tools, templates or resources would you like that you do not have? Provide up to three responses. If you do not have an answer, please leave blank and proceed to the next question

Response 1

Response 2

Response 3

On-Line Survey.

Use of F Tools and Assistance

54. Have you consulted with F on evaluation designs, evaluation Scopes of Work, or other evaluation issues?

- ☐ Yes
- ☐ No
- ☐ Don't know

56. Why were the consultations a little or not at all useful? Provide up to three responses. If you do not have an answer, please leave blank and proceed to the next question.

Response 1

Response 2

Response 3

Evaluation Community of Practice

57. Have you participated in the Evaluation Community of Practice meetings in person or through remote methods?

- ☐ Yes
- ☐ No
- ☐ Don't know

58. How often do you participate in the Evaluation Community of Practice meetings?

- ☐ All the time ☐ Not very often
☐ Most of the time ☐ Don't know
☐ Some of the time

59. How useful do you find your participation in the Evaluation Community of Practice meetings?

- ☐ Very useful ☐ Not useful
☐ Somewhat useful ☐ Don't know
☐ A little useful

60. Have you used resources or applied practices that you learned about from the Evaluation Community of Practice?

- ☐ Yes
☐ No
☐ Don't know

61. In what ways have you participated in the Evaluation Community of Practice? Check all that apply.

- ☐ Attend Community of Practice meetings in person or virtually ☐ Engage in on-line discussions with Community of Practice members
☐ Make presentations at Community of Practice meetings ☐ Engage in other forms of discussion with Community of Practice members (e.g., in-person, phone, text, Skype)
☐ Post messages at the Community of Practice website ☐ Share learning resources with Community of Practice members either by uploading them to the Community of Practice website or through other means
☐ Other (please describe)

62. What are the most useful things you have learned from your participation in the Evaluation Community of Practice? Provide up to three responses. If you do not have an answer, please leave blank and proceed to the next question.

Response 1

Response 2

Response 3

63. What would you recommend to improve Evaluation Community of Practice? Provide up to three responses. If you do not have an answer, please leave blank.

Response 1

Response 2

Response 3

Annex VI: KII and FGD Discussion Guides

KII DISCUSSION GUIDE FOR BECs

Capacity Development—Baseline Requirements of BEC Position

- What does your job as BEC entail? How much of your time do you spend on these activities?
- How would you describe your position as BEC within your Bureau/Independent Office?
- How would you describe the current evaluation capacity within your Bureau/Independent Office?
- What level of importance does your Bureau/Independent Office leadership place on evaluating the Bureau's/Office's programs? Why?
- To what extent and how does your Bureau/Independent Office leadership support the Bureau's/Office's evaluation efforts?
- How has evaluation practice within your Bureau/Independent Office changed over the last few years and particularly since the release of the new State Department Evaluation Policy in 2015? Why?
- What criteria does your Bureau/Independent Office use to select programs to be evaluated? Why?
- What challenges do you specifically as BEC and your Bureau/Independent Office in general face in designing and implementing evaluations?

Evaluation Use

- What has your experience been with the quality of evaluation reports? What are some examples of what was useful and not useful?
- To what degree would you say your Bureau/Independent Office uses evaluation findings and recommendations?
- How is your Bureau/Office using evaluation findings and recommendations?
- What would you say are the main factors that contribute to the use and non-use of evaluation findings and recommendations? Why?

Capacity Development—F Tools and Technical Assistance

- Have you taken either the Managing Evaluations or the Evaluation Designs Course sponsored by F and BP? How useful were the courses?
- To what extent have you and other people in your Bureau/Independent Office used F evaluation tools or the Evaluation Community of Practice web site for your work?
- What other evaluation technical assistance has F provided you and other people in your Bureau/Independent Office?

-How useful do you find F evaluation tools and other technical assistance to have been in improving evaluation knowledge practice within your Bureau/Independent Office? Why?

-What do you recommend to improve the usefulness of existing F evaluation tools and other technical assistance?

-Are there tools you would like to see from F that don't exist?

Capacity Development—Community of Practice

-How have you been using the State Department Evaluation Community of Practice? How useful have you found it? Why?

Capacity Development—FATAA

-We understand the new policy approved late last year incorporates fully the requirements of FATAA. To what extent is your Bureau/Independent Office prepared to meet the requirements of the new Department Evaluation Policy?

-What actions are your Bureau/Independent Office taking to meet the requirements of the new Department Evaluation Policy? (For example, are all large programs in your Bureau/Independent Office evaluated once in their lifetime or every five years or is there a plan to evaluate these programs?) Why?

-What additional support from F does your Bureau/Independent Office need to implement the new policy?

-Is there anything you would like to tell us that we haven't asked?

KII DISCUSSION GUIDE FOR DAS, A/S, AND EQUIVALENTS

- How has your bureau/independent office approached evaluation in the last 3-5 years?
- How has evaluation practice in your Bureau/Independent Office changed over the last 3-5 years, particularly since the release of the new State Department Evaluation Policy in 2015? Why?
- What have been the greatest challenges in implementing the policy?

Evaluation Use

- How does your Bureau/Independent Office use evaluation findings and recommendations?
- What would you say are the main factors that contribute to the use and non-use of evaluation findings and recommendations? Why?

Capacity Development—Baseline Requirements

- How would you describe the current evaluation capacity in your Bureau/Independent Office?
- What level of importance does your Bureau/Independent Office place on evaluating its programs? Why?
- What challenges does your Bureau/Independent Office face in doing evaluations?
- What can F do to help you meet these challenges?

Capacity Development—FATAA

- We understand the new policy approved late last year incorporates fully the requirements of FATAA. To what extent is your Bureau/Independent Office prepared to meet the requirements of the new Department Evaluation Policy?
- Is there anything you'd like to tell us that we didn't ask about?

KII DISCUSSION GUIDE FOR F AND BP

Evaluation Use

- To what degree would you say the State Department and its Bureaus/Independent Offices are using evaluation findings and recommendations?
- How are Bureaus/Independent Offices using evaluation findings and recommendations?
- What would you say are the main factors that contribute to the use and non-use of evaluation findings and recommendations? Why?

Capacity Development—Baseline Requirements

- How would you describe the current evaluation capacity within the State Department and its Bureaus/Independent Offices? Why?
- How and to what extent do you and your Office support evaluations within the Department and its Bureaus/Independent Offices?
- How has evaluation practice within the Department and its Bureaus/Independent Offices changed over the last few years, particularly since the release of the new State Department Evaluation Policy in 2015? Why?
- What challenges do the Department and its Bureaus/Independent Offices face in doing evaluations?
- What evaluation capacity gaps exist within the Department and its Bureaus/Independent Offices? What do you recommend to fill these gaps?

Capacity Development—FATAA

- We understand the new policy approved late last year incorporates fully the requirements of FATAA. Please describe how FATAA is incorporated into the policy.
- What actions are your Office taking to help the Department and its Bureaus/Independent Offices meet the requirements of the new Department Evaluation Policy?
- How do you plan to monitor implementation of the new policy?

Annex VII: List of Persons Participating in KIIs and FGDs

List of Persons Participating in KII		
Bureau/Office	Name	Position/Title
Bureau of African Affairs (AF)	Mak, Daniella	Senior M&E Advisor
Bureau of African Affairs (AF/ERA)	Thompson, Lisbeth	Sr. Strategic Planning and Budgeting Officer
Bureau of Budget and Planning (BP)	Smith, Carly	Division Director, Performance Management
	Cabell, Stephanie	Program Analyst
Bureau of Conflict and Stabilization Operations (CSO)	Gipson, Keira	Monitoring and Evaluation Lead
	Ladnier, Jason	Deputy Assistant Secretary
Bureau of Counterterrorism (CT)	Stormer, Neil	Monitoring and Evaluation Lead
Bureau of Democracy, Human Rights and Labor (DRL)	Dazzo, Giovanni	Evaluation Specialist
Bureau of East Asian and Pacific Affairs (EAP)	Cohen, Jennifer	Program Analyst
Bureau of Economic and Business Affairs (EB)	Gupman, Alexander	Sr. Coordinator for Economic Diplomacy
	Dillard, Marc	Director, Economic Policy and Public Diplomacy
European and Eurasian Affairs Office of the Coordinator for U.S. Assistance (EUR/ACE)	Wilcox, John	Director, Policy, Planning, and Performance
F	Crye, Lisa	Senior Evaluation Advisor
F/PPM	Kohari, Sarah	Performance Management Branch Chief
Office of Global Women's Issues (S/GWI)	Kuemmerle, Pamela	Senior Program Officer
	Rukundo, Sandrine	Deputy Director of Global Programs
Bureau of Intelligence and Research (INR)	Parks, Jonathan	Intelligence Operations Specialist
Bureau of International Narcotics and Law Enforcement Affairs (INL)	Thomas, Patricia	Deputy Executive Director
	Edwards, Aaron	Branch Chief, Performance Management
Bureau of International Organization Affairs (IO)	Stanley, Glenn	Effectiveness Analyst
	Hettinger, Laura	Program Analyst
	Platukis, Lauren	Effectiveness Unit Chief
Bureau of International Security and Nonproliferation (ISN/SCO)	Young, Jamie	Planning Coordinator

List of Persons Participating in KILs		
Secretary for Civilian Security, Democracy, and Human Rights J (U/S)	Lowery, Nicole	Management and Program Analyst
Bureau of Near Eastern Affairs (NEA)	Donahue, Natalie	Social Science Analyst
Bureau of Oceans and International Environmental and Scientific Affairs (OES)	Miller, Todd	Program Analyst
Bureau of Population, Refugees, and Migration (PMR)	Rowan, Maria	Program Analyst
Bureau of South and Central Asian Affairs (SCA/RA)	Holt, Thomas	Foreign Affairs Officer
Bureau of Political-Military Affairs (PM/SA)	Gobeille, Todd	Foreign Affairs Officer
S/Office of Global Partnerships (GP)	Haserodt, Jason	Program Manager
	Abu-Hamdeh, Sabrina	Senior Partnership Advisor
Office to Monitor and Combat Trafficking in Persons (J/TIP)	Allen, Karen	Foreign Affairs Officer
	Long, Abigail	Performance Measurement Specialist
Bureau of Western Hemisphere Affairs (WHA/PPC)	Ceriale, Jennifer	Deputy Director
	Crippen, John	Performance Management Officer

List of Persons Participating in FGD 1		
Bureau/Office	Name	Position/Title
Bureau of International Organization Affairs (IO)	Stanley, Glenn	Effectiveness Analyst
European and Eurasian Affairs Office of the Coordinator for U.S. Assistance (EUR)/Office of Public Diplomacy (PD)	Singh, Christine	Program Analyst
Office to Monitor and Combat Trafficking in Persons (J/TIP)	Long, Abigail	Performance Measurement Specialist

List of Persons Participating in FGD 1

Conflict and Stabilization Operations (CSO)	Hemmer, Bruce	Lead Conflict Advisor
---	---------------	-----------------------

List of Persons Participating in FGD 2

Bureau/Office	Name	Position/Title
Bureau of International Organization Affairs (IO)	Platukis, Lauren	Effectiveness Unit Chief
European and Eurasian Affairs Office of the Coordinator for U.S. Assistance (EUR)	Wilcox, Jon	Policy, Planning, and Performance
Office to Monitor and Combat Trafficking in Persons (J/TIP)	Sigmon, Jane	Senior Advisor to the Director
Conflict and Stabilization Operations (CSO)	Ladnier, Jason	Deputy Assistant Secretary
Bureau of Intelligence and Research (INR/AO)	Kuchta-Helbling, Catherine	Office Director
Bureau of Western Hemisphere Affairs (WHA/PPC)	Ceriale, Jennifer	Deputy Director

Annex VIII: Evaluation Methodology

ME&A used a mixed-methods evaluation design to collect valid and reliable data on practice and perceptions related to the implementation and use of evaluations at the DoS, the implementation of the Department's evaluation policy from 2012-2016, and DoS staffs' use of F evaluation resources. This annex provides a more detailed description than provided in the main narrative of the report about the development and implementation of these methods by the ME&A ET over the course of the evaluation.

Evaluation Design and Methods

The ME&A ET developed several methods to gather valid and reliable data to answer each of the five primary EQs posed by F in the evaluation SOW implemented during three discrete data collection phases. The first data collection phase was the Meta-Evaluation Phase, which used a meta-evaluation checklist to review the content and quality of a sample of evaluations of FA-funded programs completed across the Department between 2012 and 2016. The team analyzed these data to answer EQ 1 on the strengths and weaknesses of evaluations as shown by report content. The second data collection phase was the Performance Evaluation phase. For the second phase, the ET developed and administered an online survey targeted to BECs and other DoS staff participating in the F-supported evaluation CoP. Finally, in the third phase, the ET conducted in-person KIIs and FGDs with a purposively selected sample of BECs, bureau leaders, and other relevant DoS staff representing 25 bureaus and independent offices that have received FA funding.

The purpose of the second and third data collection phases was to generate evidence to answer EQs 2-5 on the utilization of evaluation findings and recommendations and building evaluation capacity in bureaus. The ET then analyzed the data from each data collection phase to identify findings, draw conclusions, and make recommendations to F on providing support for evaluation with the Department going forward. The team conducted the evaluation through four phases of work.

Prior to beginning data collection, the ET participated in an Inception Phase with its evaluation counterparts at F. The purpose of the Inception was to develop the needed understanding and familiarity with the DoS evaluation policy and practice so as to enable the ET to create a rigorous and comprehensive evaluation design to provide credible, evidence-based answers to the EQs.

Each of the above four evaluation phases is described in greater depth below.

Inception Phase

The Inception Phase involved several sets of activities.

Consultations with key bureau staff: To gather information and ensure that the ET had a comprehensive and accurate view of expectations for the evaluation, the ET consulted with key F staff to learn more about F's goals, activities, products, beneficiaries, and stakeholders. The ET also began discussions with F about the design, implementation, and timing for the evaluation and methods for sampling documents for the meta-evaluation and DoS staff for the online survey and qualitative data collection.

Document review: The ET conducted a desk review of documents provided by F and found through its own literature search information directly relevant to understanding the state of evaluation practice at DoS and the context for conducting the evaluation, including developments related to the DoS Evaluation Policy and FATAA. These documents included, for example, past and current versions of the DoS Evaluation Policy, FATAA, F products to support evaluation, capacity building documents, other regulatory and legal guidance on evaluations, the Department's Foreign Affairs Manual sections relevant to evaluation (FAM 301.1), assessments of DoS evaluation practice done by the OIG in 2015, and a set of completed evaluation reports from 2012-2016 that Bureaus had shared with F. As part of the document review, the ET also found and reviewed examples of other meta-evaluations that had been conducted on similar topics. A complete list of these documents is provided in Annex IV.

Development of evaluation design and all evaluation instruments: The ET used the initial consultations with F and document review to develop an evaluation design—including sampling plan, timeframe, deliverables, work plan, logistics, and data analysis plan—and the three primary research instruments: meta-evaluation checklist, online survey, and the KII/FGD discussion guides. The team held conversations with F, reviewed DoS evaluation materials, and reviewed meta-evaluations of foreign assistance programming by USAID, the Government Accountability Office (GAO), and the DoS Office of the Inspector General (OIG) to help develop a review checklist and coding for responses for the meta-evaluation along with meta-evaluation methodology for this aspect of the evaluation. All aspects of the evaluation design were thoroughly vetted between the ET and its counterparts at F and were approved by F prior to implementation.

Meta-Evaluation Phase

Developing the Meta-Evaluation Checklist

To evaluate the quality of State Department evaluation reports, the ET reviewed past efforts to assess the quality of report content at USAID and the DoS as well as for PEPFAR projects. After reviewing these past efforts, the ET decided to develop its own meta-evaluation checklist that, while drawing on the past efforts for ideas and content, focused on specific requirements or guidelines in the DoS Evaluation Policy and the official Evaluation Guidance document produced by F in November 2016 to assist bureaus in implementing the Evaluation Policy.

The ET developed an initial checklist using Excel after a round of internal discussions to identify key quality indicators and methods of measuring them. After developing an initial draft, the ET shared the draft with F for discussion and refinement. Through a process of iterative drafting, discussion, and refinement, the ET produced a penultimate version of the checklist preparatory to pilot testing. All five ET members then participated in two pilot tests of the checklist by independently reviewing the same two evaluation reports. ET members shared their completed checklists for these two sample reports with other team members and then held a meeting to discuss their use of the checklist so as to ensure inter-rater reliability. This test, discussion, and agreement created a coding system, which ensured accuracy, comprehensiveness, and consistency in the coding. The meeting also led to further elaboration of the checklist, the development of additional codes for aspects of evaluations, and modest revisions to the checklist.

The final checklist, included in Annex IV, records 11 descriptive characteristics in addition to 29 primary quality indicators and 15 supplementary quality indicators falling under nine categories: Executive Summary, report structure, evaluation objectives and audiences, EQ, methodology, findings, conclusions, recommendations, and annexes. Each primary indicator was scored using a dichotomous scoring system coded as a 1 if the report fulfilled the relevant quality criterion and as a 0 if it did not. Where the indicator did not apply for any reason, coders were instructed to score the indicator as not applicable (NA) along with the coder's explanation for the NA score for the rating in the Explanation column.

Supplemental indicators provided additional details to the dichotomous primary indicators with more detail in the form of either numerical data (e.g., Question 3a on the actual length of the main body of the evaluation report in pages) or categorical information (e.g., Question 4a listing components not explicitly included in the evaluation report). Sub-question responses, when non-numerical, were categorized and grouped by the team to capture all possible responses.

In developing the checklist, the ET made a couple of critical decisions. First, it opted for a dichotomous scoring system as opposed to scaled scoring system (e.g., 1 through 5). Its rationale for this decision was to minimize the level of subjectivity in the scoring process. Whereas some degree of subjectivity is unavoidable, using a scaled scoring system and multiple coders would significantly increase its risk of

occurrence, thereby reducing inter-rater reliability and the credibility of the meta-evaluation findings. Second, with regards to the evaluation methodology described in each report, the ET elected to limit its assessment to whether a reasonably complete description of the evaluation methodology was included in the report and not to attempt to assess the quality of the evaluation methodology. While ET members participating in the report coding each had substantial experience conducting evaluations, the team recognized that a good deal of debate persists on what constitutes sufficiently rigorous evaluation methods and attempting to litigate this debate during the meta-evaluation might introduce more problems for the ET and controversy about its scoring system than was worth the benefit for undertaking this task.

Selecting the Meta-Evaluation Report Sample

For the meta-evaluation, the ET evaluated the 72 evaluation reports, which were winnowed down from the original 124 reports provided by F during the Inception Phase. The team began with a sample of 124 reports from F that were identified as evaluations. The ET reviewed these 124 reports to verify that they were evaluation reports. In this first review, the team identified 26 files as duplicates and found that seven files were summaries of evaluations rather than full reports. The ET sent this revised list of 91 reports to F along with an explanation. F determined that the ET should analyze these reports for the meta-evaluation. At this time, F also provided the ET a link to an F online evaluation repository. The ET compared the reports at this site to the sample of 91 evaluations and identified five additional evaluations, bringing the total number of evaluations for the meta-evaluation to 96 evaluation reports.

However, when the ET conducted a second review of the files in the sample in preparation for using the checklist and codes, it realized that 19 of these 96 reports were not evaluations or were evaluations that were not comparable to the rest of the reports in the sample. The team thus excluded these 19 reports from the sample, leaving the ET with 77 evaluations for the meta-evaluation review. Excluded reports included five multi-donor Multilateral Organization Performance Assessment Network (MOPAN) evaluations, three synthesis reports, eight summaries instead of full evaluations, two quarterly reports, and one resource guide.

F disputed the ET's exclusion of eight reports from the sample, which it added back to the sample, leaving the team with 85 reports as it began meta-evaluation review. In applying the checklist to these 85 reports, ET members determined that 13 of these files were either not evaluation reports or were reports that could not be analyzed using the checklist. These included five non-evaluations, a PowerPoint presentation, five evaluation summaries, a synthesis report, and a quarterly report. Excluding these 13 left the ET with a final sample of 72 evaluation reports for the meta-evaluation.

Conducting the Meta-Evaluation

The three ET members participating in the report coding process next undertook to code the 72 evaluations over a two-week period using the meta-evaluation checklist, which consisted of 29 binary (Yes/No) primary quality indicators, falling under nine quality categories, and 15 supplemental indicators added to the checklist to add additional information to certain primary indicators for analytical purposes. In the course of coding evaluations, questions on the analysis and coding that arose were discussed and addressed as a team. This ensured that the ET continued to address all questions and issues consistently across coders. Coders inputted the results of their analysis into an Excel database, and at the conclusion of their document review, checked their entries and cleaned the data to enable quantitative analysis. Finally, the non-coder members of the ET checked the database and reviewed it for validity and reliability.

Recoding Selected Checklist Indicators

After conducting the meta-evaluation and reviewing preliminary results from the online survey, the ET realized that some of the meta-evaluation results were not matching the preliminary survey findings. A number of comments from the survey, for example, stated that some evaluation reports did not present useful information and were sometime difficult to read. Other survey respondents, moreover, did not find recommendations to be specific and actionable. However, when the team analyzed the data from the meta-evaluation, the binary nature of the scoring sheet only allowed for a yes or no answer. Consequently, nuances in the readability of a report were not portrayed in the original meta-evaluation findings. In other words, almost all the reports were readable, and recommendations were almost uniformly specific and actionable.

To understand this mismatch in the findings, the ET decided to undertake a second analysis of all the evaluation reports. The objective of this second review was to look at the nuance behind four specific indicators—report readability, recommendations follow from findings and conclusions, recommendations are specific, and recommendations are actionable—and score these on an ascending 5-point Likert scale based on the idea that this analysis might expose more nuanced differences between the reports and thus provide a better picture of report quality. So as to maximize inter-rater reliability, a single team member was assigned to undertake this re-coding exercise.

Performance Evaluation Phase

The purpose of the Performance Evaluation Phase was to collect valid and reliable data to answer EQs 2-5. The data collected were then used to evaluate 1) whether and how evaluation results are used by bureaus and what obstacles exist to their use and 2) bureau capacity and the efficacy of the Department's technical assistance to assist in the implementation of its evaluation policy. The performance evaluation phase used the following combination of quantitative and qualitative data collection methods: online staff survey, KIIs, and FGDs with DoS staff.

Online Staff Survey

The ET developed an initial set of questions to be used for an electronic survey of DoS staff engaged in evaluation during the development of the evaluation design and work plan. The survey was designed to ask Department staff about their experience with and perceptions of evaluation at State. The survey included questions to assess their views on evaluation utilization and evaluation capacity within their bureaus and offices. After creating a draft survey questionnaire during the Inception Phase, the ET revised the questionnaire based on preliminary findings from the meta-evaluation phase. After discussions with F, the ET finalized the online staff survey instrument for F approval, after which it went live on Survey Monkey.

The sampling frame for the survey was the list of the Department's Evaluation CoP. F had generated a list of staff members, including their positions, bureaus, and Department electronic mail addresses as part of their management of the CoP. The CoP list included 435 Department staff that have participated or indicated an interest in participating in the CoP. To encourage participation, F sent an initial e-mail invitation to all members of the CoP mailing list explaining the purpose for the survey and requesting their participation. After a week's time, F sent a follow-up email to all names on the CoP mailing list.

The approved work plan allotted two weeks to administer the online survey, but by the end of the second week, the response rate was below what had been envisioned, and after discussions between the ET and F, it was decided to extend the survey for an additional two weeks, at which point F sent a second follow-up email to the CoP members on the mailing list. In the end, 66 persons from 25 DoS bureaus and independent offices responded to the online survey out of the 435 names on the CoP mailing list for a response rate of 15.2 percent.

The survey response rate was probably adversely affected by the timing of the survey, which occurred near or during the traditional holiday season. The initial invitation email went out on December 11, and the original time allotted to complete the survey was from December 11 to December 22. However, and as discussed above, the decision was made to extend the survey an additional two weeks through January 5. The timing of the survey was based on the original start date of the evaluation and the contracting process, which set the times for completion of different evaluation tasks and deliverables. The ET discussed with F the possibility of changing the timing of the online survey, or extending it further beyond the holiday season, but it was decided that doing either of these options would disrupt the timeframe for the evaluation and require significant changes to the existing work plan. Thus, it was agreed to extend the survey the additional week beyond the original two weeks but to then take the responses received by that date and use those for the subsequent analysis.

The online staff survey instrument included 54 discrete questions with a combination of dichotomous (Yes/No), scaled (5-point Likert scale), and open-ended questions on respondent demographic characteristics, information about their experience with evaluation within their bureaus, their views on evaluation in their bureaus and the Department as a whole, their experience with F and use of F evaluation resources, and their priorities going forward in terms of evaluation. Respondents completed the survey in an average of 12 minutes and 10 seconds.

KIIs and FGDs

During the Inception Phase, the ET developed a set of draft KII and FGD discussion guides, which the team subsequently revised based on meta-evaluation and online survey results and via additional consultations with F. The ET created four guides, including three KII discussion guides for senior F and BP managers, bureau A/S and DAS, and BECs and other bureau staff, and one FGD discussion guide. The discussion guides included both a set of standardized questions, to allow the aggregation of information across different participant categories, and a set of customized questions, to allow the investigation of issues relevant to the specific participant category.

Over the period of January 13-28, 2018, the ET conducted 28 KIIs with 38 DoS staff and bureau leaders and two FGDs, including one FGD with seven BECs and one FGD with six higher-level bureau leaders, representing in total 25 DoS bureaus and independent offices. KIIs were designed to be in-depth, semi-structured interviews conducted with Department staff engaged in evaluations via face-to-face methods. Most KIIs were planned and held with individuals. Some KIIs were held with small groups of 2-3 key informants that worked on evaluations together in their bureaus or offices. KIIs were designed to take approximately one hour. KII questions were designed to solicit information from informants directly relevant to answering EQs 2-5. For each KII, one team member took the lead in asking and following up on questions with the informant or informants, while another team member took detailed notes on each KII. All ET members wrote up detailed interview notes at the conclusion of the KII. These notes were then shared among team members to make sure that as much information as possible was collected from key informants and used by the team.

FGDs were designed to be in-depth, semi-structured, and moderated discussions conducted face-to-face among groups of approximately six DoS staff with either similar levels of knowledge of evaluation or similar levels of responsibility with respect to evaluations. FGD methods were designed to encourage participants to discuss and debate questions among themselves, elaborating on and sharpening points of agreement and disagreement among participants. A team member facilitated the discussion to stimulate discussion among participants, while other team members took and wrote up detailed notes on the FGDs.

In selecting persons to participate in the KIIs and FGDs, the ET worked closely with its counterparts at F to construct a reasonably representative “purposive” sample of key informants. Recognizing at the outset

that interest in and capacity for conducting evaluations varied widely across DoS bureaus and independent offices, and given that the ET had a limited timeframe to conduct its face-to-face interviews that would not allow a fully representative sample of relevant key informants, it placed emphasis on selecting key informants from bureaus and independent offices at different points along the interest and capacity spectrum so as to ensure that a sufficiently diverse set of experiences and viewpoints was represented in the KIIs and FGDs. The ET also sought to ensure that it spoke with key informants both at the staff and leadership levels (e.g., BECs and DASs) with the bureaus. Finally, in selecting key informants for the KIIs and FGDs, the ET sought to identify persons with actual knowledge and experience with evaluation within the context of their bureau and DoS more broadly so as to ensure that they were capable of speaking knowledgeably about the issues of interest and answering the primary evaluation questions.

Data Analysis Phase

During the data analysis phase, the ET analyzed the quantitative and qualitative data collected during the meta-evaluation and performance evaluation focusing on the themes of quality, utilization, and capacity using the data analysis methods described below.

Quantitative Data Analysis

The meta-evaluation and online survey collected numerical, categorical, or ordinal data amenable to quantitative analysis. Even the open-ended questions in the online survey were later coded and converted into categorical data. Prior to analyzing the data, the ET conducted quality and consistency checks of the data. Once satisfied that the data was ready for analysis, the ET formatted the data in Excel, after which it converted the Excel data into a set of SPSS data files for analysis and analyzed the data using the following procedures:

- Calculated frequencies and descriptives for each indicator or variable describing the frequency of different responses and, where appropriate, the averages and other central tendencies (e.g., mean and mode).
- Calculated the same frequencies and descriptives for each meta-evaluation indicator for the four time periods of 2012-2013, 2014, 2015, and 2016 and for other report characteristics (e.g., type of evaluation, region, number of countries, and coder) and for each survey question for different respondent characteristics (e.g., BEC vs. non-BEC, years of evaluation experience, functional vs. regional bureau, and former evaluation training).
- Conducted Chi Square tests of association for both meta-evaluation and survey data so as to determine whether results (e.g., evaluation report quality, support for evaluations, use of evaluations, familiarity with the DoS Evaluation Policy, etc.) are associated with different report or respondent characteristics.

Qualitative Data Analysis

To analyze the qualitative data collected via document review, KIIs, and FGDs, the ET conducted a content analysis of the relevant documents and interview notes using the Atlas.ti qualitative analysis software. Content analysis is a well-established, credible, and systematic technique for analyzing qualitative data that compresses words from the text into fewer content categories based on explicit rules of coding. The ET then used Atlas.ti to identify and clarify patterns in the data among the codes. Content analysis allowed the ET to draw inferences from the qualitative data by objectively and systematically identifying specific themes and sub-themes within the data and assessing their relative importance in answering the EQs supported by key quotations and examples from individual documents, KIIs, or FGDs. Atlas.ti also helped the ET to understand where evidence was contradictory and whether any differences in the data were related to different factors or respondent characteristics.

Report Writing Phase

Findings, Conclusions, and Recommendations

At the conclusion of the data analysis, the ET produced files that presented the data analysis for the data collected through each of the five methods used by the team: document review, meta-evaluation, online survey, KIIs, and FGDs. Team members reviewed these written files and identified patterns relevant to the EQs. ET members next reviewed the accumulated findings, particularly findings that were confirmed by more than one source of data or method, to reach conclusions on the five primary EQs. Finally, team members wrote recommendations that followed from these conclusions. The ET shared the individual-generated findings, conclusions, and recommendations and discussed them to come to agreement on the most important, relevant, and well-founded findings, the conclusions that followed from the analysis of these findings, and the recommendations that followed from these conclusions.

Report Writing

Individual team members assumed responsibility for undertaking different aspects of the data analysis and contributing narrative sections of the evaluation report. Draft sections were shared among the team and discussed in conference calls and via email. Edited and revised versions were shared and discussed a second time before being added to the evaluation report. The ET synthesized the findings, conclusions, and recommendations from each member through this process to reach a set of agreed-on conclusions and recommendations from the accumulated findings. Next, the ET submitted a complete draft evaluation to ME&A for internal review, checking, quality control, and formatting, after which it submitted the draft evaluation report to F.

Annex IX: Online Survey Results

Survey Question	Number	Percentage
Are you a Foreign Service, Civil Service, or other type of staff member?		
Foreign Service	4	6.1
Civil Service	47	71.2
Contractor	12	19.7
Other	2	3.0
How many years have you worked at the State Department?	66	100.0
0-5	28	42.4
6-10	14	21.2
> 10	24	36.4
How many years have you worked in your current position?	66	
0-5	53	80.3
6-10	10	15.2
> 10	3	4.5
	66	
Are you a Bureau Evaluation Coordinator?		
Yes	30	45.5
No	35	53.0
Don't Know	1	1.5
	66	
Does your position (job) description include evaluation as part of the responsibilities of your position?		
Yes	49	74.2
No	14	21.3
Don't Know	3	4.5
	66	
Have you previously received formal evaluation training?		
Yes	49	74.2
No	17	25.8
	66	
How many years of evaluation experience do you have?		
1<	7	12.1
1-2	13	22.4
3-5	20	34.5
6-10	12	20.7
>10	6	10.3
	58	
How important do you think it is to evaluate your Bureau's/Office's programs, projects, or activities?		
Very Important	54	83.1
Somewhat Important	8	12.3
Not Very Important	2	3.1
Not At All Important	0	0.0
Don't Know	1	1.5
	65	
How familiar are you with the evaluations that your Bureau/Independent Office has conducted of its programs, projects, or activities?		
Very Familiar	34	52.3
Somewhat Familiar	24	36.9

Survey Question	Number	Percentage
Not Very Familiar	5	7.7
Not At All Familiar	1	1.5
Don't Know	1	1.5
	65	
Inform your Bureau's/Independent Office's strategy		
Great Extent	2	3.3
Good Extent	14	23.3
Some Extent	34	56.7
No Extent	9	15.0
Don't Know	1	1.7
	60	
Inform the design of a new program, project, or activity		
Great Extent	8	13.3
Good Extent	18	30.0
Some Extent	30	50.0
No Extent	3	5.0
Don't Know	1	1.7
	60	
Improve program, project, or activity implementation		
Great Extent	8	13.3
Good Extent	24	40.0
Some Extent	26	43.3
No Extent	1	1.7
Don't Know	1	1.7
	60	
Contribute to your Bureau's/Independent Office's learning		
Great Extent	10	16.9
Good Extent	16	27.1
Some Extent	28	47.5
No Extent	5	8.5
Don't Know	0	0.0
	59	
Support change management in your Bureau/Independent Office		
Great Extent	4	6.8
Good Extent	8	13.6
Some Extent	25	42.4
No Extent	14	23.7
Don't Know	8	13.6
	59	
Question basic assumptions about your Bureau's/Independent Office's policies or practices		
Great Extent	3	5.1
Good Extent	14	23.7
Some Extent	24	40.7
No Extent	11	18.6
Don't Know	7	11.9
	59	

Survey Question	Number	Percentage
Overall, to what extent do you believe that your Bureau/Independent Office has used evaluation findings, conclusions, and recommendations?		
Great Extent	3	5.0
Good Extent	18	30.0
Some Extent	37	61.7
No Extent	1	1.7
Don't Know	1	1.7
	60	
Failure to link the evaluation to Bureau/Independent Office or regional strategy		
Great Extent	2	3.4
Good Extent	11	18.6
Some Extent	14	23.7
No Extent	23	39.0
Don't Know	9	15.3
	59	
Lack of clear management support for the evaluation		
Great Extent	8	13.6
Good Extent	8	13.6
Some Extent	17	28.8
No Extent	19	32.2
Don't Know	7	11.9
	59	
Low awareness of the evaluation among Bureau/Independent Office staff		
Great Extent	12	20.7
Good Extent	10	17.2
Some Extent	19	32.8
No Extent	11	19.0
Don't Know	6	10.3
	58	
Low interest in the evaluation among Bureau/Independent Office staff		
Great Extent	9	15.5
Good Extent	9	15.5
Some Extent	24	41.4
No Extent	12	20.7
Don't Know	4	6.9
	58	
Failure to involve stakeholders in the evaluation process		
Great Extent	3	5.2
Good Extent	8	13.8
Some Extent	19	32.8
No Extent	20	34.5
Don't Know	8	13.8
	58	
Evaluation design or methodology that is overly complex or rigorous		
Great Extent	0	0.0
Good Extent	4	6.9
Some Extent	13	22.4

Survey Question	Number	Percentage
No Extent	33	56.9
Don't Know	8	13.8
	58	
Evaluation design or methodology that is insufficiently complex or rigorous		
Great Extent	4	6.9
Good Extent	4	6.9
Some Extent	12	20.7
No Extent	27	46.6
Don't Know	11	19.0
	58	
Evaluation reports that are too long or otherwise hard to read		
Great Extent	4	6.9
Good Extent	7	12.1
Some Extent	20	34.5
No Extent	19	32.8
Don't Know	8	13.8
	58	
Evaluation reports with unclear findings and conclusions		
Great Extent	6	10.3
Good Extent	4	6.9
Some Extent	22	37.9
No Extent	17	29.3
Don't Know	9	15.5
	57	
Evaluation reports with recommendations that are not useful or actionable		
Great Extent	5	8.8
Good Extent	7	12.3
Some Extent	24	42.1
No Extent	15	26.3
Don't Know	6	10.5
	58	
Inadequate dissemination of evaluation findings, conclusions, or recommendations		
Great Extent	9	15.5
Good Extent	6	10.3
Some Extent	22	37.9
No Extent	17	29.3
Don't Know	4	6.9
	58	
Evaluation results that come too late to inform decisions		
Great Extent	3	5.2
Good Extent	7	12.1
Some Extent	25	43.1
No Extent	18	31.0
Don't Know	5	8.6
	58	
How familiar are you with the State Department Evaluation Policy?		
Very Familiar	30	50.0

Survey Question	Number	Percentage
Somewhat Familiar	25	41.7
Not Very Familiar	3	5.0
Not At All Familiar	2	3.3
Don't Know	0	0.0
	60	
In your opinion, to what extent does the management of your Bureau/Independent Office support implementation of the Department Evaluation Policy?		
A Great Deal	18	31.0
A Good Deal	19	32.8
Some	18	31.0
Not At All	1	1.7
Don't Know	2	3.4
	58	
To what extent do you think that the Department Evaluation Policy applies to your Bureau/Independent Office?		
Completely	34	58.6
Partially	15	25.9
Some	8	13.8
Not At All	0	0.0
Don't Know	1	1.7
	58	
Overall, to what extent do you think your Bureau/Independent Office has met the requirements of the Department Evaluation Policy?		
Completely	24	42.1
Partially	23	40.4
Some	6	10.5
Not At All	0	0.0
Don't Know	4	7.0
	57	
Does your Bureau/Independent Office have a Bureau Evaluation Plan?		
Yes	36	60.0
No	7	11.7
Don't Know	17	28.3
	60	
Does your Bureau/Independent Office have a Bureau Evaluation Coordinator?		
Yes	52	86.7
No	2	3.3
Don't Know	6	10.0
How often have you consulted with your Bureau Evaluation Coordinator on evaluation issues? If you are a Bureau Evaluation Coordinator, please mark "Not applicable."		
Often	8	30.8
Some	6	23.1
A Little	4	15.4
Not At All	6	23.1

Survey Question	Number	Percentage
Don't Know	2	7.7
	26	
Overall, how useful have your consultations with your Bureau Evaluation Coordinator been? If you are a Bureau Evaluation Coordinator, please mark "Not applicable."		
Very Useful	10	45.5
Somewhat Useful	2	9.1
A Little Useful	4	18.2
Not At All Useful	5	22.7
Don't know	1	4.5
	22	
Are you familiar with the evaluation tools and templates developed by F or found at the Evaluation Community of Practice website?		
Yes	50	83.3
No	8	13.3
Don't Know	2	3.3
	60	
Writing Researchable EQs/Evaluation Design matrix template		
Very Useful	13	33.3
Somewhat Useful	4	35.9
A Little Useful	7	17.9
Not At All Useful	1	2.6
Don't know	4	10.3
	39	
SOW toolkit		
Very Useful	9	23.7
Somewhat Useful	14	36.8
A Little Useful	10	26.3
Not At All Useful	2	5.3
Don't know	3	7.9
	38	
Dissemination summary template		
Very Useful	6	16.2
Somewhat Useful	9	24.3
A Little Useful	6	16.2
Not At All Useful	1	2.7
Don't know	15	40.5
	37	
Suggestions for Reviewing Foreign Assistance Evaluation Reports		
Very Useful	4	10.5
Somewhat Useful	10	26.3
A Little Useful	8	21.1
Not At All Useful	2	5.3
Don't know	14	36.8
	38	
Low Cost or No Cost Evaluations for Activities and Small Projects		
Very Useful	5	13.2
Somewhat Useful	7	18.4

Survey Question	Number	Percentage
A Little Useful	6	15.8
Not At All Useful	14	10.5
Don't know	6	42.1
	38	
State of Evaluation Newsletters		
Very Useful	11	23.9
Somewhat Useful	10	21.7
A Little Useful	12	26.1
Not At All Useful	6	13.0
Don't know	7	15.2
	46	
Online F DOS resources		
Very Useful	10	24.4
Somewhat Useful	14	34.1
A Little Useful	8	19.5
Not At All Useful	1	2.4
Don't know	8	19.5
	41	
Evaluation WIKI		
Very Useful	2	5.1
Somewhat Useful	7	17.9
A Little Useful	3	7.7
Not At All Useful	2	5.1
Don't know	25	64.1
	39	
Independent Government Cost Estimate template (IGCE)		
Very Useful	11	26.8
Somewhat Useful	12	29.3
A Little Useful	8	19.5
Not At All Useful	2	4.9
Don't know	8	19.5
	41	
Have you consulted with F on evaluation designs, evaluation SOWs, or other evaluation issues?		
Yes	36	61.0
No	23	39.0
Don't Know	0	0.0
	59	
How useful were your consultations with F?		
Very Useful	22	59.5
Somewhat Useful	10	27.0
A Little Useful	4	10.8
Not At All Useful	1	2.7
Don't know	0	0.0
	37	
Have you participated in the Evaluation Community of Practice meetings in person or through remote methods?		
Yes	45	75.0

Survey Question	Number	Percentage
No	15	25.0
Don't Know	0	0.0
	60	
How often do you participate in the Evaluation CoP meetings?		
All the Time	5	11.1
Most of the Time	14	31.1
Some of the Time	13	28.9
Not Very Often	13	28.9
Don't Know	0	0.0
How useful do you find your participation in the Evaluation CoP meetings?	45	
Very Useful	10	22.2
Somewhat Useful	19	42.2
A Little Useful	10	22.2
Not At All Useful	3	6.7
Don't know	3	6.7
Have you used resources or applied practices that you learned about from the Evaluation CoP?	45	
Yes	27	60.0
No	14	31.1
Don't Know	4	8.9
	45	
In what ways have you participated in the Evaluation CoP?		
Attend CoP meetings	42	62.7
Present at CoP meetings	16	23.9
Post on CoP website	5	7.5
Engage online	2	3.0
Engage in other ways	19	28.4
Share resources	7	10.4
Other	0	0.0
Functional vs. Regional Bureau		
Technical	50	76.1
Regional	16	23.9
	66	
Bureau/Independent Office		
International Narcotics and Law Enforcement Affairs (INL)	13	19.4
Africa Affairs (AF)	7	10.4
Foreign Assistance (F)	4	6.0
International Security and Nonproliferation (ISN)	4	6.0
Political-Military Affairs (PM)	4	6.0
International Organization Affairs (IO)	3	4.5
Trafficking in Persons (J/TIP)	3	4.5
Conflict Stabilization and Operations (CSO)	2	3.0
Counter Terrorism (CT)	2	3.0
Democracy, Human Rights and Labour (DRL)	2	3.0
Economic and Business Affairs (EB)	2	3.0
Europe and Eurasian Affairs (EUR)	2	3.0
Near Eastern Affairs (NEA)	2	3.0

Survey Question	Number	Percentage
Oceans and International Environmental and Scientific Affairs (OES)	2	3.0
Population, Refugees and Migration (PRM)	2	3.0
South and Central Asian Affairs (SCA)	2	3.0
Western Hemisphere Affairs (WHA)	2	3.0
Arms Control, Verification and Compliance (AVC)	1	1.5
East Asian and Pacific Affairs (EAP)	1	1.5
Energy Resources (ENR)	1	1.5
Global Partnerships (S/GPI)	1	1.5
Global Women's Issues (S/GWI)	1	1.5
Intelligence and Research (INR)	1	1.5
International Information Programs (IIP)	1	1.5
J Front Office	1	1.5
Office of the Undersecretary for Civilian Security, Democracy and Human Rights	1	1.5

Annex X: Relationships between Survey Responses and Different Respondent Categories

The ET conducted a series of Chi Square tests to determine whether answers to survey questions manifested statistically significant differences between respondents belonging to different respondent groups. Specifically, the Chi Square tests looked at whether responses differed for respondents belonging to the following groups: 1) BECs vs. non-BECs; 2) respondents with evaluation responsibility vs. respondents without evaluation responsibility; 3) respondents with evaluation training vs. respondents without evaluation training; 4) respondents with more evaluation experience vs. respondents with less evaluation experience; and 5) respondents more familiar with evaluations done by their bureau vs. respondents less familiar with evaluations done by their bureau. The table below shows the Chi Square statistic and level of statistical significance for those relationships found to be statistically significant. An explanation of the results is found below the table.

Survey Question	BEC	Evaluation Responsibility	Previous Evaluation Training	Evaluation Experience	Familiar with Evaluations
Importance of evaluating bureau programs	5.17***	10.41*	-	-	-
Familiarity with bureau evaluations	6.39**	-	12.39*	-	-
Evaluation use: Inform bureau strategy	-	7.83**	-	-	-
Hinder: Failure to link evaluation to bureau strategy	7.89**	6.28**	-	-	-
Hinder: Lack of management support	7.81**	-	9.21**	-	-
Hinder: Low awareness of bureau staff	7.03***	-	-	-	-
Hinder: Failure to involve stakeholders	15.08*	-	-	-	-
Hinder: Evaluation design insufficiently rigorous	10.44*	-	-	11.35***	-
Hinder: Recommendations not specific or actionable	-	8.47**	-	-	-
Hinder: Evaluation results too late	7.80**	10.22**	-	-	-
Familiarity with DoS evaluation policy	-	-	15.53*	-	18.01*
Bureau management supports DoS evaluation policy	8.47**	-	-	-	-
Bureau in compliance with DoS evaluation policy	-	-	-	-	9.42*
Familiarity with F evaluation tools	-	-	10.71*	-	3.23***
Consulted with F	3.09***	-	-	-	5.33**
How useful were your consultations with F?	-	10.53*	-	-	-
Participated in CoP	3.71**	-	7.00*	-	5.40**
Usefulness of CoP participation	-	-	-	11.00***	-

*Statistically significant at the .01 level

**Statistically significant at the .05 level

***Statistically significant at the .10 level

BEC vs. Non-BEC: Survey respondents who are BECs are more likely to acknowledge the importance of doing evaluations, are more familiar with the evaluations done by their bureaus, are more likely to say that their bureau management supports the DoS evaluation policy, consult more often with F, and participate more in the Evaluation CoP. In contrast, they are less likely to say that the following hinder the use of evaluations within their bureaus: failing to link evaluations to bureau strategy, lack of management support, low awareness of evaluations by bureau staff, failure to involve stakeholders in the evaluation process, and evaluations that come too late to inform decisions.

Evaluation responsibility vs. No evaluation responsibility: Survey respondents who have responsibility for

evaluations within their bureau are more likely to acknowledge the importance of doing evaluations, more likely to say their bureau uses evaluations to inform strategy, more likely to say that failing to link evaluations to bureau strategy hinders evaluation use, less likely to say that poor evaluation recommendations or evaluation results that come too late hinder evaluation use, and find consultations with F more useful.

Previous evaluation training vs. No previous evaluation training: Respondents with previous evaluation training are more familiar with the evaluations done by their bureau, more likely to say that the lack of management support hinders evaluation use, more familiar with the DoS evaluation policy, more familiar with the F evaluation tools, and participate more in the evaluation CoP.


More evaluation experience vs. Less evaluation experience: Survey respondents with less evaluation experience are more likely to say that insufficiently rigorous evaluations hinder evaluation use and find CoP participation less useful.

More familiar with bureau evaluations vs. Less familiar with bureau evaluations: Survey respondents who are more familiar with the evaluations conducted by their bureau are more optimistic about their bureau management's support of the DoS evaluation policy and their bureau's compliance with the evaluation policy, are more familiar with the F evaluation tools, consult with F more, and participate in the evaluation CoP more.

Annex XI: Disclosure of any Conflicts of Interest

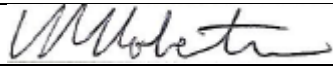
Name	Gary Woller
Title	Senior Evaluation Specialist
Organization	ME&A
Evaluation Position?	X Team Leader <input type="checkbox"/> Team member
Evaluation Award Number (or RFTOP or other appropriate instrument number)	Department of State (DoS) Performance Management and Evaluation Services IDIQ Functional Area 6 D17PC00465, Task Order D17PS00713
DoS Project(s) Evaluated (Include project name(s), implementer name(s) and award number(s), if applicable)	Examination of Foreign Assistance Evaluation Efforts at the Department of State: A Combined Performance and Meta-Evaluation
I have real or potential conflict of interest to disclose.	<input type="checkbox"/> Yes X No
If yes answered above, I disclose the following facts: <i>Real or potential conflicts of interest may include, but are not limited to:</i> <ol style="list-style-type: none"> 1. Close family member who is an employee of the DoS operating unit managing the project(s) being evaluated or the implementing organization(s) whose project(s) are being evaluated. 2. Financial interest that is direct, or is significant though indirect, in the implementing organization(s) whose projects are being evaluated or in the outcome of the evaluation. 3. Current or previous direct or significant though indirect experience with the project(s) being evaluated, including involvement in the project design or previous iterations of the project. 4. Current or previous work experience or seeking employment with the DoS operating unit managing the evaluation or the implementing organization(s) whose project(s) are being evaluated. 5. Current or previous work experience with an organization that may be seen as an industry competitor with the implementing organization(s) whose project(s) are being evaluated. 6. Preconceived ideas toward individuals, groups, organizations, or objectives of the particular projects and organizations being evaluated that could bias the evaluation. 	

I certify (1) that I have completed this disclosure form fully and to the best of my ability and (2) that I will update this disclosure form promptly if relevant circumstances change. If I gain access to proprietary information of other companies, then I agree to protect their information from unauthorized use or disclosure for as long as it remains proprietary and refrain from using the information for any purpose other than that for which it was furnished.

Signature	
Date	February 8, 2018

Name	Lawrence Robertson
Title	Senior Evaluation Specialist
Organization	ME&A
Evaluation Position?	<input type="checkbox"/> Team Leader X Team member
Evaluation Award Number (or RFTOP or other appropriate instrument number)	Department of State (DoS) Performance Management and Evaluation Services IDIQ Functional Area 6 D17PC00465, Task Order D17PS00713
DoS Project(s) Evaluated (Include project name(s), implementer name(s) and award number(s), if applicable)	Examination of Foreign Assistance Evaluation Efforts at the Department of State: A Combined Performance and Meta-Evaluation
I have real or potential conflict of interest to disclose.	<input type="checkbox"/> Yes X No
If yes answered above, I disclose the following facts: <i>Real or potential conflicts of interest may include, but are not limited to:</i> <ol style="list-style-type: none"> 1. Close family member who is an employee of the DoS operating unit managing the project(s) being evaluated or the implementing organization(s) whose project(s) are being evaluated. 2. Financial interest that is direct, or is significant though indirect, in the implementing organization(s) whose projects are being evaluated or in the outcome of the evaluation. 3. Current or previous direct or significant though indirect experience with the project(s) being evaluated, including involvement in the project design or previous iterations of the project. 4. Current or previous work experience or seeking employment with the DoS operating unit managing the evaluation or the implementing organization(s) whose project(s) are being evaluated. 5. Current or previous work experience with an organization that may be seen as an industry competitor with the implementing organization(s) whose project(s) are being evaluated. 6. Preconceived ideas toward individuals, groups, organizations, or objectives of the particular projects and organizations being evaluated that could bias the evaluation. 	

I certify (1) that I have completed this disclosure form fully and to the best of my ability and (2) that I will update this disclosure form promptly if relevant circumstances change. If I gain access to proprietary information of other companies, then I agree to protect their information from unauthorized use or disclosure for as long as it remains proprietary and refrain from using the information for any purpose other than that for which it was furnished.


Signature	
Date	February 8, 2018

Name	Christopher Coffman
Title	Project Manager / Research Analyst
Organization	ME&A
Evaluation Position?	<input type="checkbox"/> Team Leader X Team member
Evaluation Award Number (or RFTOP or other appropriate instrument number)	D17PD01299
DoS Project(s) Evaluated (Include project name(s), implementer name(s) and award number(s), if applicable)	N/A
I have real or potential conflict of interest to disclose.	<input type="checkbox"/> Yes X No
<p>If yes answered above, I disclose the following facts:</p> <p><i>Real or potential conflicts of interest may include, but are not limited to:</i></p> <ol style="list-style-type: none"> 1. Close family member who is an employee of the DoS operating unit managing the project(s) being evaluated or the implementing organization(s) whose project(s) are being evaluated. 2. Financial interest that is direct, or is significant though indirect, in the implementing organization(s) whose projects are being evaluated or in the outcome of the evaluation. 3. Current or previous direct or significant though indirect experience with the project(s) being evaluated, including involvement in the project design or previous iterations of the project. 4. Current or previous work experience or seeking employment with the DoS operating unit managing the evaluation or the implementing organization(s) whose project(s) are being evaluated. 5. Current or previous work experience with an organization that may be seen as an industry competitor with the implementing organization(s) whose project(s) are being evaluated. 6. Preconceived ideas toward individuals, groups, organizations, or objectives of the particular projects and organizations being evaluated that could bias the evaluation. 	

Signature	
Date	February 14, 2018

Name	Dereje Gebreegzabher
Title	Research Analyst
Organization	ME&A
Evaluation Position?	Team Leader X Team member
Evaluation Award Number (or RFTOP or other appropriate instrument number)	Department of State (DoS) Performance Management and Evaluation Services IDIQ Functional Area 6 D17PC00465, Task Order D17PS00713
DoS Project(s) Evaluated (Include project name(s), implementer name(s) and award number(s), if applicable)	Examination of Foreign Assistance Evaluation Efforts at the Department of State: A Combined Performance and Meta-Evaluation
I have real or potential conflict of interest to disclose.	<input type="checkbox"/> Yes X No
If yes answered above, I disclose the following facts: <i>Real or potential conflicts of interest may include, but are not limited to:</i> <ol style="list-style-type: none"> 1. Close family member who is an employee of the DoS operating unit managing the project(s) being evaluated or the implementing organization(s) whose project(s) are being evaluated. 2. Financial interest that is direct, or is significant though indirect, in the implementing organization(s) whose projects are being evaluated or in the outcome of the evaluation. 3. Current or previous direct or significant though indirect experience with the project(s) being evaluated, including involvement in the project design or previous iterations of the project. 4. Current or previous work experience or seeking employment with the DoS operating unit managing the evaluation or the implementing organization(s) whose project(s) are being evaluated. 5. Current or previous work experience with an organization that may be seen as an industry competitor with the implementing organization(s) whose project(s) are being evaluated. 6. Preconceived ideas toward individuals, groups, organizations, or objectives of the particular projects and organizations being evaluated that could bias the evaluation. 	

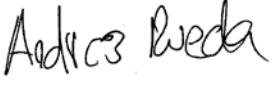
I certify (1) that I have completed this disclosure form fully and to the best of my ability and (2) that I will update this disclosure form promptly if relevant circumstances change. If I gain access to proprietary information of other companies, then I agree to protect their information from unauthorized use or disclosure for as long as it remains proprietary and refrain from using the

Signature	
Date	10-18-2017

information for any purpose other than that for which it was furnished.

Name	Andres Rueda Cardenas
Title	Research Associate
Organization	ME&A
Evaluation Position?	Team Leader X Team member
Evaluation Award Number (or RFTOP or other appropriate instrument number)	Department of State (DoS) Performance Management and Evaluation Services IDIQ Functional Area 6 D17PC00465, Task Order D17PS00713
DoS Project(s) Evaluated (Include project name(s), implementer name(s) and award number(s), if applicable)	Examination of Foreign Assistance Evaluation Efforts at the Department of State: A Combined Performance and Meta-Evaluation
I have real or potential conflict of interest to disclose.	<input type="checkbox"/> Yes X No
If yes answered above, I disclose the following facts: <i>Real or potential conflicts of interest may include, but are not limited to:</i> <ol style="list-style-type: none"> 1. Close family member who is an employee of the DoS operating unit managing the project(s) being evaluated or the implementing organization(s) whose project(s) are being evaluated. 2. Financial interest that is direct, or is significant though indirect, in the implementing organization(s) whose projects are being evaluated or in the outcome of the evaluation. 3. Current or previous direct or significant though indirect experience with the project(s) being evaluated, including involvement in the project design or previous iterations of the project. 4. Current or previous work experience or seeking employment with the DoS operating unit managing the evaluation or the implementing organization(s) whose project(s) are being evaluated. 5. Current or previous work experience with an organization that may be seen as an industry competitor with the implementing organization(s) whose project(s) are being evaluated. 6. Preconceived ideas toward individuals, groups, organizations, or objectives of the particular projects and organizations being evaluated that could bias the evaluation. 	

I certify (1) that I have completed this disclosure form fully and to the best of my ability and (2) that I will update this disclosure form promptly if relevant circumstances change. If I gain access to proprietary information of other companies, then I agree to protect their information from unauthorized use or disclosure for as long as it remains proprietary and refrain from using the information for any purpose other than that for which it was furnished.

Signature	
Date	Feb 16, 2018

U.S. Department of State
2201 C Street NW
Washington, DC 20520