



March 2017

FOREIGN ASSISTANCE

Agencies Can Improve the Quality and Dissemination of Program Evaluations

Why GAO Did This Study

The U.S. government plans to spend approximately \$35 billion on foreign assistance in 2017. Evaluation is an essential tool for U.S. agencies to assess and improve the results of their programs. Government-wide guidance emphasizes the importance of evaluation, and the Foreign Aid Transparency and Accountability Act of 2016 requires the President to establish guidelines for conducting evaluations. However, evaluations can be challenging to conduct. GAO has previously reported on challenges in the design, implementation, and dissemination of the evaluations of individual foreign assistance programs.

GAO was asked to review foreign aid evaluations across multiple agencies. This report examines the (1) quality, (2) cost, and (3) dissemination of foreign aid program evaluations. GAO assessed a representative sample of 173 fiscal year 2015 evaluations for programs at the six agencies providing the largest amounts of U.S. foreign aid—USAID, State, MCC, HHS’s Centers for Disease Control and Prevention under the President’s Emergency Plan for AIDS Relief, USDA’s Foreign Agricultural Service, and DOD’s Global Train and Equip program—against leading evaluation quality criteria; analyzed cost and contract documents; and reviewed agency websites and dissemination procedures.

What GAO Recommends

GAO recommends that each of the six agencies develop a plan to improve the quality of its evaluations and that HHS, MCC, State, and USDA improve their procedures and planning for disseminating evaluation reports.

The agencies concurred with our recommendations.

View [GAO-17-316](#). For more information, contact Jessica Farb at (202) 512-6991 or farbj@gao.gov.

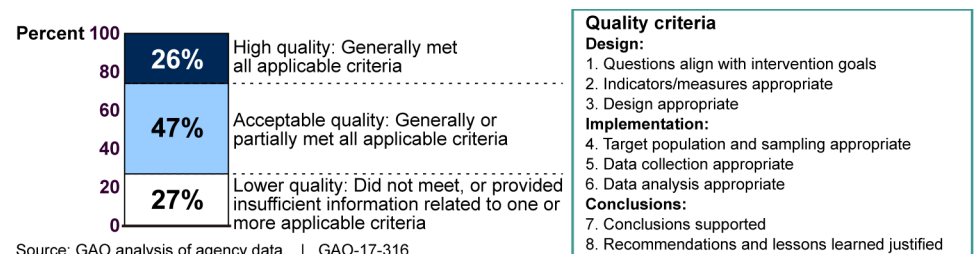
FOREIGN ASSISTANCE

Agencies Can Improve the Quality and Dissemination of Program Evaluations

What GAO Found

An estimated 73 percent of evaluations completed in fiscal year 2015 by the six U.S. agencies GAO reviewed generally or partially addressed all of the quality criteria GAO identified for evaluation design, implementation, and conclusions (see fig.). Agencies met some elements of the criteria more often than others. For example, approximately 90 percent of all evaluations addressed questions that are generally aligned with program goals and were thus able to provide useful information about program results. About 40 percent of evaluations did not use generally appropriate sampling, data collection, or analysis methods. Although implementing evaluations overseas poses significant methodological challenges, GAO identified opportunities for each agency to improve evaluation quality and thereby strengthen its ability to manage aid funds more effectively based on results.

Estimated Percentage of Foreign Assistance Evaluations Meeting Evaluation Quality Criteria



Source: GAO analysis of agency data. | GAO-17-316

Note: The confidence intervals for our estimates of the quality of agency evaluations according to these categories did not exceed ± 8 percent.

Evaluation costs ranged widely and were sometimes difficult to determine, but the majority of evaluations GAO examined cost less than \$200,000. Millennium Challenge Corporation (MCC) evaluations had a median cost of about \$269,000, while median costs for the U.S. Agency for International Development (USAID), the U.S. Department of Agriculture (USDA), and the Department of State (State) ranged from about \$88,000 to about \$178,000. GAO was unable to identify the specific costs for the Department of Defense (DOD) and Department of Health and Human Services (HHS) evaluations. High-quality evaluations tend to be more costly, but some well-designed lower-cost evaluations also met all quality criteria. Other factors related to evaluation costs include the evaluation’s choice of methodology, its duration, and its location.

Agencies generally posted and distributed evaluations for the use of internal and external stakeholders. However, shortfalls in some agency efforts may limit the evaluations’ usefulness.

- **Public posting.** USDA has not developed procedures for reviewing and preparing its evaluations for public posting, but the other agencies posted nonsensitive reports on a public website.
- **Timeliness.** Some HHS reports and more than half of MCC reports were posted a year or more after completion.
- **Dissemination planning.** State does not currently have a policy requiring a plan that identifies potential users and the means of dissemination.

Contents

Letter		1
	Background	4
	Most Foreign Aid Evaluations Were of High or Acceptable Quality Overall, though Quality Varied by Criterion and Agency	8
	Foreign Aid Evaluation Costs Range Widely and Are Influenced by Methodology, Location, and Evaluation Quality	20
	Selected Agencies' Evaluations Are Generally Available Online, but Some Agencies Can Improve Dissemination	25
	Conclusions	31
	Recommendations for Executive Action	31
	Agency Comments and Our Evaluation	32
Appendix I	Objectives, Scope, and Methodology	36
Appendix II	Evaluation Review Data, by Agency	50
Appendix III	Comments from the Department of Defense	66
Appendix IV	Comments from the Department of Health and Human Services	69
Appendix V	Comments from the Millennium Challenge Corporation	71
Appendix VI	Comments from the Department of State	74
Appendix VII	Comments from the U.S. Agency for International Development	76
Appendix VIII	Comments from the U.S. Department of Agriculture	79

Tables

Table 1: Estimated Percentages of Agency Evaluations Generally or Partially Meeting Applicable Quality Criteria or Not Meeting One or More Criteria	11
Table 2: Extent to Which Evaluations Generally Met Quality Criteria, by Agency	18
Table 3: Fiscal Year 2015 Foreign Assistance Evaluation Costs, by Agency and Evaluation Type	22
Table 4: Fiscal Year 2015 MCC, State, USAID, and USDA Foreign Assistance Evaluation Costs, by Quality Category	24
Table 5: Assessment of Six U.S. Agencies' Use of Six Practices for Effective Dissemination of Foreign Aid Evaluations for Fiscal Year 2015	26
Table 6: Agency Foreign Assistance Evaluation Study Population and Sampling	38
Table 7: Number of Quality Criteria Generally Met, by Quality Category	44
Table 8: Number of Evaluations in Cost Sample, by Agency	45
Table 9: Quality and Characteristics of the Design of Foreign Assistance Evaluations	51
Table 10: Quality and Characteristics of the Implementation of Foreign Assistance Evaluations	52
Table 11: Quality and Characteristics of the Conclusions of Foreign Assistance Evaluations	53
Table 12: Quality and Characteristics of the Design of HHS PEPFAR Evaluations	53
Table 13: Quality and Characteristics of the Implementation of HHS PEPFAR Evaluations	54
Table 14: Quality and Characteristics of the Conclusions of HHS PEPFAR Evaluations	55
Table 15: Quality and Characteristics of the Design of MCC Evaluations	56
Table 16: Quality and Characteristics of the Implementation of MCC Evaluations	57
Table 17: Quality and Characteristics of the Conclusions of MCC Evaluations	58
Table 18: Quality and Characteristics of the Design of State Evaluations	58

Table 19: Quality and Characteristics of the Implementation of State Evaluations	59
Table 20: Quality and Characteristics of the Conclusions of State Evaluations	60
Table 21: Quality and Characteristics of the Design of USAID Evaluations	61
Table 22: Quality and Characteristics of the Implementation of USAID Evaluations	62
Table 23: Quality and Characteristics of the Conclusions of USAID Evaluations	63
Table 24: Quality and Characteristics of the Design of USDA Evaluations	63
Table 25: Quality and Characteristics of the Implementation of USDA Evaluations	64
Table 26: Quality and Characteristics of the Conclusions of USDA Evaluations	65

Figures

Figure 1: Estimated Percentages of Foreign Assistance Evaluations Meeting Evaluation Quality Criteria	13
Figure 2: Distribution of Costs for Fiscal Year 2015 Millennium Challenge Corporation, Department of State, U.S. Agency for International Development, and Department of Agriculture Evaluations Reviewed, by Evaluation Type	21

Abbreviations

ADS	Automated Directives System
AEA	American Evaluation Association
CDC	Centers for Disease Control and Prevention
DAC	Development Assistance Committee
DCI	data collection instrument
DOD	Department of Defense
FAS	Foreign Agricultural Service
FPDS-NG	Federal Procurement Data System – Next Generation
GSA	General Services Administration
GT&E	Global Train and Equip
HHS	Department of Health and Human Services
MCC	Millennium Challenge Corporation
M&E	monitoring and evaluation
NDAA	National Defense Authorization Act
OECD DAC	Organization for Economic Co-operation and Development, Development Assistance Committee
OMB	Office of Management and Budget
PEPFAR	President's Emergency Plan for AIDS Relief
State	Department of State
USAID	U.S. Agency for International Development
USDA	U.S. Department of Agriculture

This is a work of the U.S. government and is not subject to copyright protection in the United States. The published product may be reproduced and distributed in its entirety without further permission from GAO. However, because this work may contain copyrighted images or other material, permission from the copyright holder may be necessary if you wish to reproduce this material separately.



March 3, 2017

Congressional Requesters

The U.S. government plans to spend approximately \$35 billion on foreign assistance in 2017 to improve the lives and health of millions living in poverty, support democracy, enhance global security, and achieve other U.S. foreign policy goals. For U.S. agencies that provide foreign assistance, evaluations are essential to assess and help improve program results.¹ Preparing and disseminating high-quality evaluations helps agencies and their implementing partners assess their program results, adjust program designs, and make evidence-based decisions about the use of their resources. Both the 2010 GPRA Modernization Act² and the 2010 Presidential Policy Directive on Global Development Policy³ called for an increased focus on evaluations of agency programs. In addition, in July 2016, the Foreign Aid Transparency and Accountability Act of 2016⁴ required the President to set forth guidelines for the establishment of measurable goals, performance metrics, and monitoring and evaluation (M&E) plans for U.S. foreign assistance within 18 months of its enactment. In recent years, foreign assistance agencies have adopted or updated their guidance on evaluations. However, prior GAO work has identified challenges in the design, implementation, and dissemination of evaluations of individual foreign assistance programs.

We were asked to review U.S. agencies' evaluation of foreign assistance programs. Focusing on evaluations completed in fiscal year 2015 by the six agencies that administer the largest amounts of U.S. foreign assistance—the Department of Defense (DOD), the Department of Health and Human Services (HHS), the Millennium Challenge Corporation

¹Evaluations are systematic studies conducted periodically or on an ad hoc basis to assess how well a program is working and to learn the benefits of a program or how to improve it. See GAO, *Performance Measurement and Evaluation: Definitions and Relationships*, [GAO-11-646SP](#) (Washington, D.C.: May 2011).

²Pub. L. No. 111-352, 124 Stat. 3866 (2011). The GPRA Modernization Act of 2010 aims to ensure that agencies use performance information in decision making and holds them accountable for achieving results and improving government performance.

³Presidential Policy Directive-6 *U.S. Global Development Policy* (Washington, D.C.: Sept. 22, 2010).

⁴Foreign Aid Transparency and Accountability Act of 2016, P.L. 114-191, July 15, 2016.

(MCC), the Department of State (State), the U.S. Department of Agriculture (USDA), and the U.S. Agency for International Development (USAID)⁵—this report examines (1) the extent to which foreign assistance program evaluations met key evaluation quality criteria, (2) the costs of the agencies' evaluations and factors that affect these costs, and (3) the extent to which the agencies ensure the dissemination of evaluation reports within the agency and to the public.

To identify the six agencies that administer the largest amounts of foreign assistance, we reviewed obligations data that the agencies reported to USAID's U.S. Overseas Loans and Grants database for fiscal years 2008 through 2012. To identify evaluations completed in fiscal year 2015, we requested that each agency provide a list of all foreign aid evaluation reports completed in that year. We did not separately review agency files to identify if agencies had additional evaluations beyond those listed in the registries. We performed an initial review of the evaluation lists and documents provided by agencies and excluded some documents from our review because they were incomplete, were not evaluation reports, or were not completed in fiscal year 2015.

To address our first objective, we reviewed all State, DOD, and MCC evaluation reports completed in fiscal year 2015. We reviewed representative samples of USAID, USDA, and HHS evaluation reports to create estimates about the population of all evaluation reports at the sampled agencies.⁶ We reviewed the selected evaluations against eight criteria for high-quality evaluations related to the appropriateness of design, data collection methods, and analysis and the extent of support for conclusions and any recommendations. We developed these criteria on the basis of a review of federal, international, and evaluation

⁵The order in which the six agencies are listed is alphabetical and does not reflect the amounts of foreign assistance they administer. For DOD, HHS, and USDA—we reviewed evaluations covering specific programs: the President's Emergency Plan for AIDS Relief (PEPFAR) programs implemented by HHS's Centers for Disease Control and Prevention (CDC); the Food for Progress and McGovern-Dole food aid programs administered by USDA's Foreign Agricultural Service (FAS); and the Global Train and Equip (GT&E) program, which focuses on security assistance and cooperation, administered by DOD. The remaining three agencies—MCC, State, and USAID—focus exclusively on foreign affairs or foreign assistance.

⁶All percentage estimates for aggregated results from our review have margins of error at the 95 percent confidence level of plus or minus 8 percentage points or less, unless otherwise noted, and all percentage estimates for individual agencies from our review have margins of error at the 95 percent confidence level of plus or minus 11 percentage points or less, unless otherwise noted.

organization guidance and our prior reports (See app. II for the criteria and data by agency for our review). We assessed each agency's evaluations as "generally," "partially," or "not at all" meeting each criterion; we also rated some evaluations as providing insufficient information to make an assessment. In addition to assessing evaluation quality, we also collected information about the characteristics of each evaluation, such as the location of the study and its methodology.

To address our second objective, we reviewed contract documents, invoices, and related documents to determine the cumulative cost of final evaluations conducted by an outside evaluator. We defined the cumulative costs as the cost of conducting the final evaluation and any related activities that informed the final evaluation, such as a midterm or baseline evaluation. In some cases, we were unable to determine an evaluation's precise cost if it was procured under a contract that covered additional activities. In these cases, we approximated the cost on the basis of estimates provided by the agency or contractor. We did not determine the cost of evaluations prepared by agency staff because agencies did not separately track these costs. To identify factors that affect the costs of foreign aid evaluations, we analyzed the cost of MCC, State, USDA, and USAID evaluations in relation to the data we collected on these evaluations' quality and other characteristics. We report only limited data on the cost of DOD's GT&E and HHS's PEPFAR evaluations because the evaluation contracts or implementing partner agreements did not separately track evaluation costs, and we concluded that the available estimates were too limited to include in our statistical analysis.

To address our third objective, we identified leading practices for the dissemination of evaluation findings. We identified these leading practices using federal guidance that encourages the timely public posting of agency information on a searchable website, as well as plans and additional efforts to actively disseminate agency information. In addition to the federal guidance, we also used the American Evaluation Association's (AEA) *An Evaluation Roadmap for a More Effective Government* (AEA Roadmap),⁷ as well as some other nonfederal sources that also cite timely public posting, dissemination planning, and additional active efforts to disseminate results as important communication tools for evaluations. We then reviewed each agency's evaluation policies to identify their

⁷American Evaluation Association, "An Evaluation Roadmap for a More Effective Government" (October 2013), accessed November 10, 2016, <http://www.eval.org/d/do/472>.

requirements for dissemination of evaluation reports and interviewed cognizant officials. We compared agency policies and practices with the leading practices we identified. We reviewed agency websites to determine whether evaluation reports were posted online and examined each agency website to determine whether it provided a search engine that could be used to locate evaluations.

See appendix I for a more detailed discussion of our scope and methodology.

We conducted this performance audit from October 2015 to March 2017 in accordance with generally accepted government auditing standards. Those standards require that we plan and perform the audit to obtain sufficient, appropriate evidence to provide a reasonable basis for our findings and conclusions based on our audit objectives. We believe that the evidence obtained provides a reasonable basis for our findings and conclusions based on our audit objectives.

Background

The six agencies whose evaluations we reviewed focus on foreign assistance to varying degrees. DOD, HHS, and USDA provide foreign assistance as part of their larger portfolios of programs, while MCC, State, and USAID focus exclusively on foreign affairs or foreign assistance.

- DOD's GT&E program provides training, equipment, and small-scale military construction activities to partner nations to build their capacity and enable them to conduct counterterrorism operations or to support ongoing allied or coalition military or stability operations that benefit the national security interests of the United States.⁸

⁸DOD's GT&E program was originally authorized as a temporary program under Section 1206 of the fiscal year 2006 National Defense Authorization Act. The fiscal year 2015 National Defense Authorization Act, Pub. L. No. 113-291, permanently authorized the GT&E program as Section 2282. The fiscal year 2017 National Defense Authorization Act enacted a new Chapter 16 within Title 10 of the U.S. Code that will contain various authorities related to defense security cooperation and, specifically, a new authority to build the capacity of foreign security forces to be codified at 10 U.S.C. § 333. The act repeals Section 2282 270 days after its enactment. See Pub. L. No. 114-328 § 1241.

-
- HHS's CDC implements a portion of the President's Emergency Plan for AIDS Relief (PEPFAR) programs under the direction of State's Office of the U.S. Global AIDS Coordinator and Health Diplomacy.⁹
 - MCC, a U.S. government corporation, provides aid to developing countries that have demonstrated a commitment to ruling justly, encouraging economic freedom, and investing in people. MCC supplies this assistance to eligible countries primarily through 5-year compacts with the goal of reducing poverty by stimulating economic growth.
 - State, the lead U.S. foreign affairs agency, implements programs that provide, for example, counternarcotics assistance; refugee assistance; and support for democracy, governance, and human rights.
 - USAID, the lead U.S. foreign assistance agency, implements programs intended to both further America's interests and improve lives in the developing world. USAID's broad portfolio includes programs that address democracy and human rights, water and sanitation, food security, education, poverty, the environment, global health, and other areas.
 - USDA's Foreign Agricultural Service (FAS) administers two nonemergency food aid programs: (1) The Food for Progress program supports agricultural value chain development, expanding revenue and production capacity, and increasing incomes in food-insecure countries; (2) The McGovern-Dole International Food for Education and Child Nutrition program supports education and nutrition for schoolchildren, particularly girls, expectant mothers, and infants.

Agency Evaluation Guidance

Each of the six agencies has adopted evaluation guidance for the programs included in our review.¹⁰

⁹We previously reported on PEPFAR evaluation quality and guidance in May 2012. See GAO, *President's Emergency Plan for AIDS Relief: Agencies Can Enhance Evaluation Quality, Planning, and Dissemination*, [GAO-12-673](#) (Washington, D.C.: May 31, 2012).

¹⁰We have previously reported on these six foreign assistance agencies' M&E policies and the consistency of these policies with leading practices. We found that, with the exception of DOD's, the agencies' foreign assistance M&E policies that we reviewed generally addressed the leading practices we identified. See GAO, *Foreign Assistance: Selected Agencies' Monitoring and Evaluation Policies Generally Address Leading Practices*, [GAO-16-861R](#) (Washington, D.C.: Sept. 27, 2016).

-
- DOD's November 2012 Section 1206 Assessment Handbook serves as a guide to evaluation planners and implementers for conducting evaluations of DOD's GT&E programs as required by federal law. The fiscal year 2012 National Defense Authorization Act (NDAA) required DOD, no later than 90 days after the end of each fiscal year, to submit to Congress a report including an assessment of the effectiveness of GT&E programs conducted that fiscal year in building the capacity of the recipient foreign country. The fiscal year 2015 NDAA maintained this requirement through 2020.¹¹ DOD did not have agency-wide evaluation guidance for security cooperation at the time we performed our review but issued such guidance in January 2017.¹²
 - For PEPFAR programs, including those implemented by HHS, State's Office of the U.S. Global AIDS Coordinator and Health Diplomacy issued the PEPFAR Evaluation Standards of Practice in January 2014 and an updated version (version 2) in September 2015.¹³
 - MCC's May 2012 *Policy for Monitoring and Evaluation of Compacts and Threshold Programs* requires that compact M&E plans identify and describe the evaluations that will be conducted, key evaluation questions and methodologies, and data collection strategies.
 - State issued its current evaluation policy and an additional guidance document for evaluations in January 2015 and issued a revised and updated version of the guidance in January 2016.
 - USAID lays out its evaluation policies in its Automated Directives System (ADS). USAID issued a fully revised ADS 201, addressing evaluation guidance, planning, and implementation, in September 2016.
 - USDA's FAS evaluations are guided by its May 2013 *Monitoring and Evaluation Policy*, which requires both interim and final program evaluations.

¹¹In April 2016, we reported on GT&E program management and reporting, including the results of its evaluations. See GAO, *Counterterrorism: DOD Should Enhance Management of and Reporting on Its Global Train and Equip Program*, [GAO-16-368](#) (Washington, D.C.: Apr. 18, 2016).

¹²DOD Instruction 5132.14, *Assessment, Monitoring, and Evaluation Policy for the Security Cooperation Enterprise*, January 13, 2017, available at <http://www.dtic.mil/whs/directives>.

¹³In March 2016, CDC issued its operationalization of the PEPFAR Evaluation Standards of Practice and added agency-specific evaluation requirements. An update to the CDC operationalization guidance was published to CDC staff in January 2017.

Agency Evaluation Procurement and Cost Tracking

With the exception of HHS, the agencies we selected for our review generally rely on outside contractors to conduct evaluations. DOD, MCC, State, and USAID directly contract for third-party evaluation services. The HHS PEPFAR evaluations we reviewed were prepared (1) by CDC staff using existing program data; (2) by an implementing partner as part of the partner's cooperative agreement; or (3) in one instance, under a separate agreement. USDA implementing partners procured the USDA evaluations whose costs we reviewed.

The six agencies track evaluation costs to varying extents. DOD, MCC, and State procured the evaluations we reviewed through centrally managed contracts, and cost information for these evaluations was available through the program or agency evaluation office. HHS's PEPFAR, USDA, and USAID evaluations were often procured and managed at the country, mission, or implementing partner level. Cost information was not centrally available and could be obtained only from each mission or implementing partner.

Evaluation Types, Timing, and Methods

Foreign assistance evaluations may vary in type, timing, and method. Two common types of evaluation are the following:

- *Performance evaluations* assess the extent to which a program is operating as was intended or the extent to which it achieves its outcome-oriented objectives. Performance evaluations often judge program effectiveness against criteria, such as progress against baselines, whether program goals were met, or whether expected targets were met.
- *Net impact evaluations* assess the net effect of a program by comparing program outcomes with an estimate of what would have happened in the program's absence. Net impact evaluations use a variety of experimental and quasi-experimental designs, including randomized methods in which participants are assigned to separate control or treatment groups to isolate the program's effect. Net impact evaluations have more complex methodologies than the other evaluation types.

Agencies may conduct evaluations during or after the completion of a program. Interim or midterm evaluations are conducted while a program is in progress, and final evaluations are conducted after the program ends. Baseline evaluations are also sometimes conducted before a program begins as a basis for determining any effects of the program.

Evaluations may use one or more methods to produce their results. For example, evaluations may use random or nonrandom sampling from the target population to select cases for inclusion in the study. Evaluations may also use one or more methods to collect data on the chosen indicators and measures—for example, structured or unstructured interviews, focus groups, surveys, direct observations, or collection and analysis of existing data. Each of these methods has potential benefits and limitations that an evaluator must consider in assessing the evaluation’s evidence as a basis for its conclusions and recommendations.

Most Foreign Aid Evaluations Were of High or Acceptable Quality Overall, though Quality Varied by Criterion and Agency

Overall, about three quarters of all 2015 foreign aid evaluations from the six agencies we reviewed generally or partially met the quality criteria we identified. The remaining evaluations did not meet one or more of these criteria or provided insufficient information. While we generally found that evaluations met quality criteria related to design, implementation, and conclusions, we more often found limitations in implementation—including sampling methods, data collection, and analysis. In addition, we found that the independence of evaluators was not always clearly evident. While the quality of evaluations varied by agency, we identified shortcomings at all six of the selected agencies that could limit evaluation reliability and usefulness.

About Three-Quarters of Agencies' Evaluations Showed High or Acceptable Quality Overall

By reviewing policies of federal agencies,¹⁴ international organizations,¹⁵ and evaluation organizations,¹⁶ and our prior reporting,¹⁷ we identified common characteristics of high-quality evaluations, from which we developed eight criteria for assessing evaluation quality. These quality criteria are associated with the (1) design, (2) implementation, and (3) conclusions of an evaluation, as follows. (See app. I for a full description of how we developed our evaluation criteria.)

Design

- Evaluation questions are aligned with program goals.
- Performance indicators are appropriate for measuring progress against program goals.
- Design is appropriate for answering the evaluation questions.

Implementation

- Target population and sampling method are appropriate, given the scope and nature of the evaluation questions.
- Data collection is appropriate for answering the evaluation questions.
- Data analysis is appropriate to answer the evaluation questions.

Conclusions

- Conclusions are supported by the available evidence.

¹⁴We reviewed agency evaluation guidance at all of the selected agencies except DOD in preparing our criteria for evaluation quality. DOD did not have applicable agency-wide evaluation criteria as we undertook our review but issued such guidance in January 2017. See [GAO-16-861R](#).

¹⁵United Kingdom Department for International Development, *International Development Evaluation Policy*, May 2013, accessed January 30, 2017, https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/204119/DFID-Evaluation-Policy-2013.pdf.

¹⁶AEA Roadmap.

¹⁷See [GAO-12-673](#); GAO, *Designing Evaluations: 2012 Revision*, [GAO-12-208G](#) (Washington, D.C.: January 2012); GAO, *Water and Sanitation Assistance: USAID Has Increased Strategic Focus but Should Improve Monitoring*, [GAO-16-81](#) (Washington, D.C. Oct. 6, 2015); and GAO, *International Cash-Based Food Assistance: USAID Has Established Processes to Monitor Cash and Voucher Projects, but Data Limitations Impede Evaluation*, [GAO-16-819](#) (Washington, D.C.: Sept. 20, 2016).

-
- Recommendations and lessons learned are justified by the available evidence.

Based on an assessment of agency evaluations against these criteria, we rated 73 percent of all evaluations as high quality (26 percent) or acceptable quality (47 percent), because they generally or partially met all applicable quality criteria. We rated the remaining 27 percent as lower quality because they either did not meet or did not provide sufficient information related to at least one applicable criterion. These evaluations may not provide sufficiently reliable evidence to inform agency program and budget decisions. Overall, we encountered more instances when evaluations did not provide sufficient information about a certain criterion than instances when evaluations did not meet a quality criterion at all.¹⁸ Table 1 summarizes our observations about the quality of evaluations at the six selected agencies in our review.¹⁹

¹⁸For example, for the quality criterion related to the appropriate definition of the target population and use of sampling techniques for the study questions, we found that about 11 percent of the evaluations did not provide sufficient information compared to 1 percent of the evaluations that did not meet this criterion at all.

¹⁹A higher percentage of evaluations that were designed to assess programs' net impacts were of higher quality than those designed to assess performance, primarily because of differences in the implementation of their design. See app. I for more information about the implementation challenges by evaluation type.

Table 1: Estimated Percentages of Agency Evaluations Generally or Partially Meeting Applicable Quality Criteria or Not Meeting One or More Criteria

	Percentage of evaluations						
	All agencies	DOD	HHS	MCC	State	USAID	USDA
High quality: Generally met all applicable criteria.	26	0	35	44	4	26	21
Acceptable quality but could be improved: Generally or partially met all applicable criteria but did not generally meet all.	47	50	38	44	48	49	48
Lower quality: Did not meet, or provided insufficient information related to, one or more applicable criteria.	27	50	26	13	48	26	30
Number of evaluations reviewed	173	4	34	16	23	63	33

Legend: DOD = Department of Defense Global Train and Equip program, HHS = Department of Health and Human Services Centers for Disease Control & Prevention–President’s Emergency Plan for AIDS Relief, MCC = Millennium Challenge Corporation, State = Department of State, USAID = U.S. Agency for International Development, USDA = U.S. Department of Agriculture Foreign Agricultural Service food assistance programs.

Source: GAO analysis of fiscal year 2015 agency evaluation reports. | GAO-17-316

Notes: Percentages for all agencies combined and for USAID are weighted to reflect the evaluations in the population that were not selected for the sample. The confidence interval for all six agencies did not exceed ±8 percent for the seven criteria that were based on the full sample of 173 evaluations and did not exceed ±9 percent for the variable that relied on the 161 evaluations that had recommendations and lessons learned. The confidence intervals for the estimates for USAID, USDA, and HHS did not exceed ±11 percent except for the variable that relied on the 161 evaluations that had recommendations and lessons learned, where it did not exceed ±14 percent.

We assessed 161 evaluations against all eight quality criteria. We assessed the 12 evaluations that did not include recommendations or lessons learned against the remaining seven applicable criteria.

Columns may not sum to 100 percent because of rounding.

Examples of High-Quality Evaluations

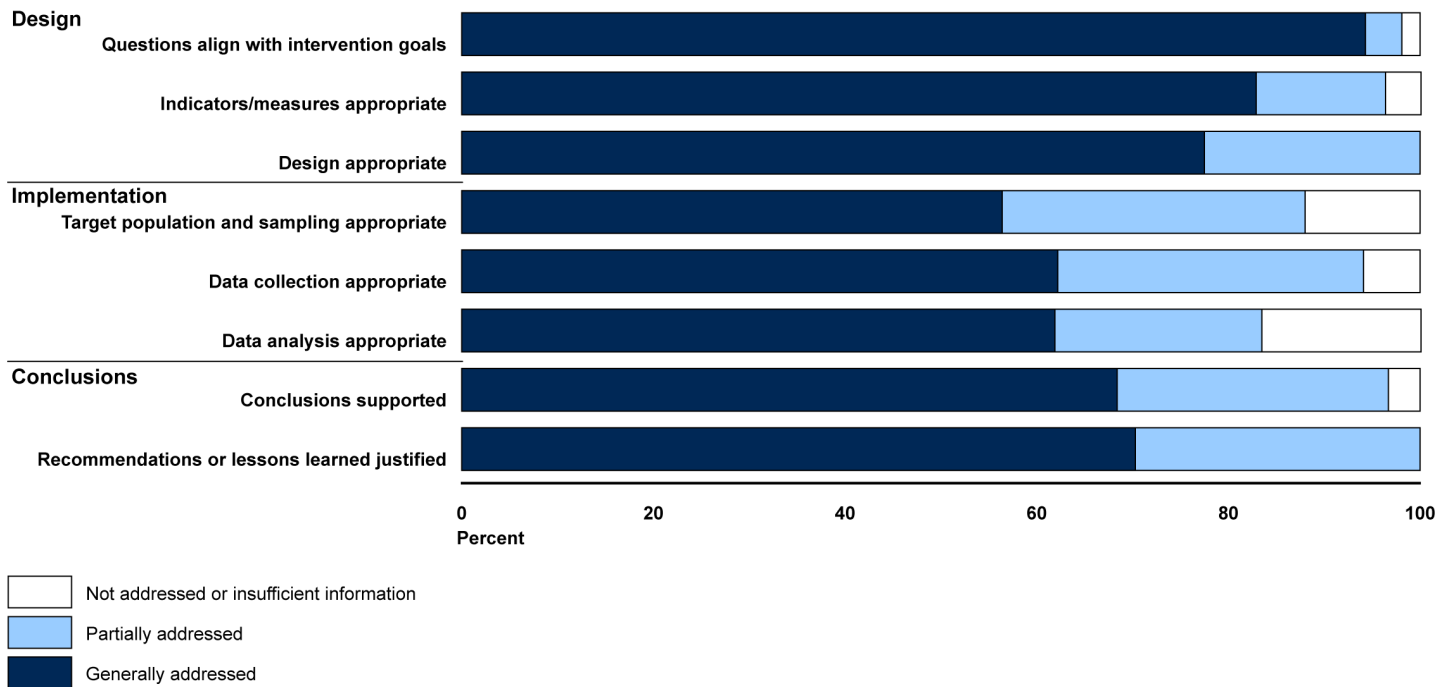
- *An HHS evaluation that focused on the implementation and use of a particular laboratory test and its associated technology and training.* The study design, target population, and sampling were all generally appropriate, and the indicators focused on implementation concerns rather than impacts. The evaluation adequately stated conclusions and causal inferences that were tied to the evidence and made explicit cautions about the effects of possible underutilization of the new technology.
- *An MCC evaluation of a rural road rehabilitation project intended to reduce transport costs over several years.* This net impact evaluation included baseline data and targets, discussed analysis techniques and sensitivity analysis thoroughly, and made causal inferences in appropriate ways with the necessary caveats.
- *A State evaluation of a corrections system program in the Middle East.* This performance evaluation used an appropriate mix of qualitative and quantitative methods to assess increases in the knowledge and skills of corrections officers. The evaluation used random sampling from lists of current correctional officers trained and others, as well as nonrandom selection of prison sites and key informants for focus group discussions. The evaluation tried to mitigate any limitations due to conducting a study in an unsafe country environment and from the potential lack of trustworthiness from inmate self-reports. As a result, the conclusions and recommendation were carefully worded to take account of limitations and were justified by the available evidence.
- *A USAID evaluation that focused on providing a range of technical assistance to regional and governmental entities in the agricultural sector in Africa.* The evaluation design was carefully thought out given the evaluation questions addressing implementation and outcome issues. The target population and sampling methods were carefully discussed and defended, and the data collection included multiple methods: an Internet survey, interviews, and a review of other data. The report provided a clear and detailed description of steps taken in the analysis and coding of qualitative data, which supported its conclusions and recommendations.
- *A USDA midterm evaluation that aimed to assess the extent to which a program improved the quality of education and nutrition in schools in a Latin American country.* The evaluation used mixed design and data collection involving sample surveys, focus groups, and in-depth interviews to assess outcomes and processes based on a range of nutrition and literacy indicators. The evaluation included both output and outcome metrics and assessed program results against literacy and other targets set relative to baselines and project plans. The evaluation used appropriate sampling methods for both random and nonrandom sampling of schools, students, teachers, and mothers of students and drew conclusions that were supported by collected evidence.

Source: GAO analysis of fiscal year 2015 agency evaluation reports. | GAO-17-316

Criteria Related to Evaluation Implementation Posed Greatest Challenge for Agencies' Evaluations

The quality of evaluations varied by the type of quality criterion we applied. As figure 1 shows, many evaluations generally met the criteria related to the appropriateness of evaluation design, implementation, and conclusions. However, overall, evaluations generally met fewer criteria related to implementation, reflecting limitations in the way evidence was collected or analyzed.

Figure 1: Estimated Percentages of Foreign Assistance Evaluations Meeting Evaluation Quality Criteria



Source: GAO analysis of fiscal year 2015 agency evaluation reports. | GAO-17-316

Notes: We assessed 161 evaluations against all eight quality criteria. We assessed 12 evaluations that did not include recommendations or lessons learned against the remaining seven applicable criteria.

Percentages for all agencies combined and for the U.S. Agency for International Development (USAID) are weighted to reflect the evaluations in the population that were not selected for the sample. The confidence interval for all six agencies did not exceed ± 8 percent for the seven criteria that were based on the full sample of 173 evaluations and did not exceed ± 9 percent for the variable that relied on the 161 evaluations that had recommendations and lessons learned.

Percentages shown for each criterion may not sum to 100 because of rounding.

We reviewed evaluations of foreign assistance programs administered by the Department of Defense Global Train and Equip program; Department of Health and Human Services's Centers for Disease Control & Prevention—President's Emergency Plan for AIDS Relief; Millennium Challenge Corporation; Department of State; USAID; and U.S. Department of Agriculture Foreign Agricultural Service food assistance programs.

Evaluation Designs Were Generally Appropriate

A relatively high percentage of evaluations generally met each of the criteria we used to assess the alignment of the study questions with the program goals, the appropriateness of the evaluation design for the study questions, and the use of indicators for measuring progress.

- **Alignment of questions with program goals.** Evaluation questions generally aligned with one or more of the evaluated program's goals

Implementation of Sampling,
Data Collection, and Analysis
More Often Had Limitations

in more than 90 percent of the evaluations. Thus, evaluations were designed to provide useful information about program results.²⁰

- **Appropriate evaluation design for the study questions.** About 80 percent of the evaluations used a design that was generally appropriate for the study questions, and the remainder of the designs was at least partially appropriate.²¹
- **Appropriate use of indicators.** Indicators for measuring progress were generally appropriate in about 80 percent of the evaluations. As a result, successes or failures identified by these evaluations are likely to be directly relevant to assessing achievement of the evaluated programs' goals.²²

We found more limitations in the implementation of evaluations than in their design. On average, about 60 percent of evaluations generally met each of the criteria related to this aspect of quality—sampling, data collection, and analysis. Limitations we identified revealed that conducting evaluations overseas can pose challenges for evaluators. For example, travel to remote areas with safety and security concerns may limit an evaluator's ability to conduct appropriate sampling and collect primary data for the study. Also, insufficient local resources to implement certain methodologies, such as implementing survey instruments, or a lack of local administrative data on the study population may constitute additional obstacles to sampling and data collection.²³

²⁰Because the goals of some programs were broader than those addressed by the evaluation study questions, an evaluation could generally meet this criterion without addressing every program goal.

²¹Almost all (98 percent) of the evaluations assessed some outcomes, with qualitative discussions about outcomes as well as findings based on outcome metrics. About 90 percent were designed to assess processes such as program implementation, and about 20 percent were designed to compare results across groups or time periods to assess program net impact (e.g., randomized trials or other control or comparison groups, time series, or statistical controls).

²²We found that about 80 percent of the evaluations included specific output metrics and about 85 percent had outcome metrics. However, only about 50 percent of the evaluations had baselines, about 50 percent had targets, and some did not use either output or outcome metrics. Without such metrics, it is not possible to quantify the progress achieved for a measurable indicator compared with established baselines and targets.

²³See app. I for our methodology, including a discussion of the varying characteristics of the programs that the evaluations under our review covered.

Limitations in Sampling Methods

About 40 percent of evaluations had limitations in, or provided insufficient information about, their sampling methodology. If an evaluation does not clearly describe how sampling was conducted, it may raise questions about the quality of the evidence, including concerns regarding selection bias of respondents, sufficiency of the sample size for the findings, and the relevance of the evaluation's findings and conclusions for the entire study population. Sampling methods were particularly problematic or unclear in evaluations that used nonrandom sampling. For evaluations that relied primarily or exclusively on testimonial evidence, the method for selecting participants for interviews or focus group discussions was sometimes inappropriate or unclear. For example, one evaluation we reviewed relied largely on interviews but did not describe the process used for selecting participants, and it indicated that the available list of potential participants was incomplete and inaccurate. Several evaluations provided insufficient information about the target population, other than identifying them as program beneficiaries, and included no discussion of how participants were selected for interviews, focus groups, or surveys.

Limitations in Data Collection Methods

About 40 percent of the evaluations had limitations in, or provided insufficient information about, their data collection methods.²⁴ We identified a number of deficiencies in the data collection process, including a lack of documentation of data collection instruments (DCI), such as questionnaires or structured interview protocols. In cases where evidence was gathered through a DCI, some evaluations were unclear about how the instrument was designed and administered. For example, an evaluation of a program intended to increase access to mobile technologies and improve mothers' health used a survey to gather data. However, the evaluation did not provide sufficient details about the survey, such as the questionnaire itself or the sampling strategy, for the reader to be able to determine the validity and reliability of the data collected.

²⁴Evaluations used a wide range of data collection methods. Overall, about 90 percent of the evaluations used semistructured or unstructured interviews, about 80 percent used administrative program data, about 60 percent used focus groups, about 40 percent used surveys of program beneficiaries, and about 40 percent used direct observations.

In addition, about half of the evaluations did not collect baseline data from which to calculate change after the program was implemented, and about half generally did not set targets, which makes an assessment of progress toward meeting the goals of the program difficult. For example, an evaluation of two community training programs in an Asian country identified the study question but did not collect baseline data and did not establish targets. Although some baseline data may have been gathered by the implementing partner, such information was not used for comparative purposes, making it impossible to assess the net effects of the program.

Further, an estimated 60 percent of the evaluations used data collection procedures that only partially ensured the reliability of the data, or there was not sufficient information to assess data reliability. For example, an evaluation of a small business program in Latin America acknowledged numerous data quality problems, including serious attrition among the group used as a comparison group to program participants. Such limitations raise questions about the strength of the conclusions.

Limitations in Data Analysis

About 40 percent of evaluations did not demonstrate that they had conducted appropriate data analysis. These evaluations often did not specify the analysis methods for each question, such as how interview responses were analyzed.²⁵ For example, an evaluation of a program serving women living with HIV analyzed data through a content analysis but did not clearly explain the categories created for the analysis or the numbers of individual responses that fell into each category. The lack of clarity in the analysis makes it difficult for the reader to determine whether the findings from this program have broader applicability. Several evaluations relied on focus group discussions but analyzed and reported the percentages of informants expressing the stated views in ways that did not appropriately account for the potential influence of other focus group members on informants' responses. Evaluations that used some quantitative data analysis also had certain shortcomings. For example, an evaluation reported a statistically significant change from baseline but did

²⁵In addition, some evaluations did not adequately document key assumptions and did not perform any robustness checks or sensitivity analysis on the methodology used to analyze the data. For example, one evaluation relied on a key assumption about applying data from later years to earlier years to estimate the likelihood of transfusion transmission of infection, but it included no discussion of sensitivity analyses or robustness checks.

not include a discussion of the type of statistical test that supported this result.

Finally, while about 90 percent of the evaluations assessed processes such as program implementation, about half of those evaluations did not establish any criteria, such as evaluation plans, budgets, timeframes, and targets. Without such benchmarks, it is difficult to define what constituted success for the evaluated program.

Evaluation Conclusions and Recommendations Were Generally Supported

The majority of evaluations generally met each of our criteria related to conclusions. These evaluations considered the strengths and limitations of the available evidence from the evaluation's design and implementation and included conclusions that were generally supported and recommendations that were generally justified.

- **Conclusions supported by the available evidence.** About 70 percent of the evaluations had conclusions that were generally supported by the evidence, and nearly all of the evaluations had conclusions that were partially supported. This indicates that these evaluations did not reach beyond what was supported by the evidence and justified given the limitations.
- **Recommendations and lessons learned justified by the available evidence.** About 75 percent of evaluations with recommendations included evidence that generally supported the recommendations, and all evaluations with recommendations included evidence that at least partially supported them. This indicates that the collected evidence justified the follow-up steps the evaluations recommended.

Independence of Evaluators Was Not Always Clearly Documented

Our analysis found that, in addition to meeting the eight criteria to varying extents, the evaluations did not always provide documentation of the evaluator's independence and whether there were any potential conflicts of interest. In instances where an evaluation was not conducted by a third party, a statement about conflicts of interest may be especially important to forestall any potential concerns about the evaluator's impartiality. In all, about 80 percent of the agency evaluations documented that they were conducted by third-party evaluators, while about 13 percent were not conducted by a third-party evaluator, and another 6 percent did not indicate whether they were performed by a third-party evaluator. About 70 percent of HHS evaluations, about 30 percent of State evaluations, and about 40 percent of USAID evaluations included a conflict-of-interest statement, while no DOD, USDA, or MCC evaluations included such a statement. If an evaluation does not address the independence of the

evaluation organization and of individual evaluators, questions could arise about the objectivity and reliability of the evaluation’s findings.

Extent to Which Evaluations Met Each Applicable Quality Criterion Varied among Agencies

As table 2 shows, the extent to which the evaluations met each quality criterion varied among the six agencies we reviewed. While our assessments revealed strengths in each agency’s evaluations, all six agencies’ evaluations also showed shortcomings in quality that could limit the agencies’ ability to ensure the effectiveness of foreign assistance based on evaluation results.²⁶

Table 2: Extent to Which Evaluations Generally Met Quality Criteria, by Agency

Criterion	Percentage of evaluations						
	Total	DOD	HHS	MCC	State	USAID	USDA
Study questions align with the key stated goal(s) of the intervention.	94	100	85	94	96	96	94
The chosen indicators/measures are appropriate for the study objectives.	83	50	97	88	48	83	88
The evaluation design is appropriate given the study questions.	78	25	79	88	57	80	73
The target population and sampling for the evaluation are appropriate for the study questions.	56	0	62	63	43	59	48
The data collection is appropriate for the study questions.	62	0	76	69	35	63	61
The data analysis appears appropriate to the task.	62	25	74	69	48	63	52
Conclusions are supported by the available evidence.	68	25	65	75	61	73	52
Recommendations and lessons learned are supported by the available evidence.	70	33	80	79	86	65	82
Number of evaluations	173	4	34	16	23	63	33

Legend: DOD = Department of Defense—Global Train & Equip; HHS= Health and Human Services, Centers for Disease Control & Prevention—President’s Emergency Plan for AIDS Relief; MCC = Millennium Challenge Corporation; State = Department of State; USAID = U.S. Agency for International Development; USDA = U.S. Department of Agriculture—Foreign Agricultural Service.

Source: GAO analysis of fiscal year 2015 agency evaluations. | GAO-17-316

Notes: We assessed 161 evaluations with recommendations against all eight quality criteria and assessed 12 evaluations without recommendations against the remaining seven applicable criteria.

Percentages for all agencies combined and for USAID are weighted to reflect the evaluations in the population that were not selected for the sample. The confidence interval for all six agencies did not exceed ±8 percent for the seven criteria that were based on the full sample of 173 evaluations and

²⁶See app. II for additional detail on our evaluation quality findings by agency.

did not exceed ± 9 percent for the variable that relied on the 161 evaluations that had recommendations and lessons learned. The confidence intervals for the estimates for USAID, USDA, and HHS did not exceed ± 11 percent except for the variable that relied on the 161 evaluations that had recommendations and lessons learned, where it did not exceed 14 percent.

Each applicable quality criterion was generally met by a majority of HHS, MCC, and USAID evaluations. However, evaluations for all three agencies scored generally lower on the criteria related to evaluation implementation—that is, the appropriateness of the target population and sampling, data collection, and data analysis. While most HHS, MCC, and USAID evaluations used generally appropriate sampling methods, our analysis showed that overall about half of the evaluations did not use appropriate nonrandom sampling techniques. In addition, we estimate that overall only about half of the three agencies' evaluations generally used data collection methods that ensured data reliability, and only about 10 to 20 percent of USAID and HHS evaluations generally specified the key assumptions of the data analysis methods used.

DOD's GT&E program evaluations' study questions met the first quality criterion—aligning with the program's goals—but overall did not generally meet the other criteria. For example, we identified weaknesses in the implementation of the evaluations' designs in terms of target population and sampling, data collection, and analysis. In particular, some evaluations did not describe the target population and did not discuss the methods the evaluators used for their selection of the equipment items they observed or the persons they interviewed. In addition, we found limited discussion about how the data were summarized and analyzed, incomplete baseline metrics, and a lack of targets.²⁷ Without systematic selection of equipment to observe or respondents to interview, it is difficult to know whether the selections were justifiable and selected in a way that supports the intervention's objectives and the conclusions drawn. Because of these implementation weaknesses as well as a lack of discussion of study limitations, we determined that the DOD GT&E program evaluations provided only partial support for their conclusions.

²⁷DOD's fiscal year 2015 GT&E program guidance states that a baseline assessment of recipient unit capabilities should be completed prior to submission of each program proposal. In April 2016, we reported that only 34 of the 51 assessments we reviewed included completed baseline assessment sections in the proposal. We recommended that DOD take steps to ensure that documentation requested in project proposal packages is complete. See [GAO-16-368](#).

State and USDA evaluations each met more than one quality criterion about half the time or less. About half or fewer of State evaluations generally met four criteria: appropriate indicators, appropriate target population and sampling, appropriate data collection, and appropriate data analysis. Regarding the appropriateness of chosen indicators, less than a third of State evaluations had indicators with baselines or established criteria such as plans or budgets, and almost none of the evaluations had indicators with targets against which progress could be assessed. In addition, about 80 percent of State evaluations used a data collection process that did not generally ensure the reliability of the data, and about half of the State evaluations generally did not specify data analysis methods for each question and the key assumptions used in the analysis. State officials noted that their programs are often implemented rapidly in response to specific events, making it difficult to design an evaluation for the program and to gather baseline data. We estimate that overall about half of USDA evaluations had generally appropriate target population and sampling, generally appropriate data analysis, or support for conclusions.

Foreign Aid Evaluation Costs Range Widely and Are Influenced by Methodology, Location, and Evaluation Quality

Most foreign aid evaluations we reviewed cost less than \$200,000, but costs ranged widely and varied by agency and type. We identified costs for MCC, State, USAID, and USDA final evaluations but could not obtain specific cost information for DOD's GT&E and HHS's PEPFAR evaluations because these programs used procurement methods for their evaluations that did not separately track evaluation costs. Evaluation costs were related to the evaluation's methodology and location, and higher-cost evaluations tended to meet more evaluation quality criteria, though we also identified lower-cost evaluations that met all quality criteria.

Evaluation Costs Ranged Widely and Varied by Type and Agency, but Most Cost Less Than \$200,000

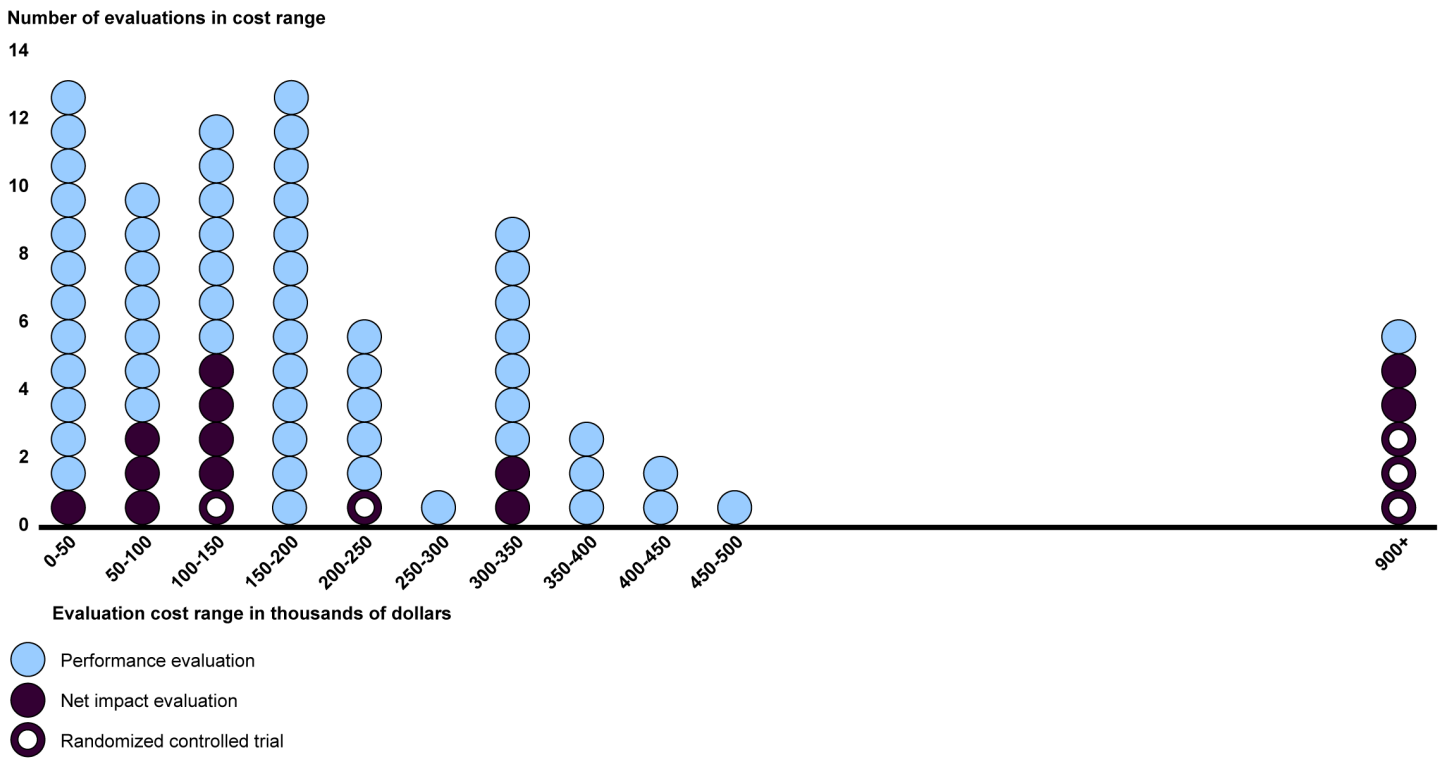
Costs for the majority of the foreign aid evaluations whose costs we reviewed were less than \$200,000, but the costs ranged widely and varied by type of evaluation and agency. Of the 76 MCC, State, USAID, and USDA evaluations, 48 cost less than \$200,000 while 6 cost more than \$900,000. The costs of net impact evaluations ranged from approximately \$36,100 to \$2.2 million, with a median of \$117,500.²⁸ The

²⁸In this report, net impact evaluations are evaluations that GAO determined to include an estimate of the net impacts of a foreign aid intervention. Individual agencies may categorize these evaluations differently. See app. I for further information on how we selected and categorized evaluations.

costs of performance evaluations ranged from \$9,600 to \$902,100, with a median cost of \$169,600.

While the median cost was higher for the performance evaluations than the net impact evaluations, the net impact evaluations had a higher average cost than the performance evaluations; five of the six evaluations that cost more than \$900,000 were net impact evaluations. Net impact evaluations that used randomized controlled trials were the most expensive evaluations in our sample, with a median cost of \$926,600 compared with \$154,700 for all other evaluations. Figure 2 shows the range of costs for the net impact and performance evaluations whose costs we reviewed.

Figure 2: Distribution of Costs for Fiscal Year 2015 Millennium Challenge Corporation, Department of State, U.S. Agency for International Development, and Department of Agriculture Evaluations Reviewed, by Evaluation Type



Source: GAO analysis of agency data. | GAO-17-316

Of the four agencies' evaluations whose costs we reviewed, MCC's evaluations had the highest median cost, at \$268,900, and USDA's evaluations had the lowest median cost, at \$87,900 (see table 3). Seven of 12 MCC evaluations cost over \$200,000, including 3 net impact evaluations that cost over \$900,000. In contrast, 8 of the 10 USDA evaluations were performance evaluations that cost less than \$200,000. Most State evaluations were performance evaluations, which were generally more expensive than performance evaluations at the other agencies. USAID costs for impact and performance evaluations both ranged widely, and USAID's net impact evaluations had a lower median cost than its performance evaluations.

Table 3: Fiscal Year 2015 Foreign Assistance Evaluation Costs, by Agency and Evaluation Type

(Dollars in thousands)

Agency/ evaluation type	Number of evaluations	Evaluation cost (dollars)			
		Average	Median	Minimum	Maximum
MCC	12	\$520.2	268.9	9.6	2,230.4
Net impact	6	\$863.8	657.1	85.0	2,230.4
Performance	6	\$176.5	107.5	9.6	457.0
State	16	\$248.6	177.7	38.3	902.1
Net impact	2	\$211.0	211.0	79.0	343.0
Performance	14	\$254.0	177.7	38.3	902.1
USAID	38	\$207.8	147.3	22.1	1,012.5
Net impact	8	\$340.2	117.2	97.1	1,012.5
Performance	30	\$172.5	159.8	22.1	407.5
USDA	10	\$129.9	87.9	26.5	401.2
Net impact	1	\$36.1	36.1	36.1	36.1
Performance	9	\$140.3	122.8	26.5	401.2

Legend: MCC = Millennium Challenge Corporation, State = Department of State, USDA = U.S. Department of Agriculture, USAID = U.S. Agency for International Development.

Sources: GAO analysis of evaluation contracts, invoices, and related documents; data from Federal Procurement Data System-Next Generation. | GAO-17-316

Note: We could not obtain specific cost information for the Department of Defense's Global Train and Equip and the Department of Health and Human Services' President's Emergency Plan for AIDS Relief evaluations.

Costs for DOD's GT&E evaluations and HHS's PEPFAR evaluations were not specifically identifiable because they were not separately tracked by the agencies, contractors, or implementing partners.

- The contract for the DOD GT&E evaluations included many activities in addition to the evaluations and was not structured to show the cost

of each activity. Additionally, according to DOD officials, neither DOD nor the contractor separately tracked the evaluation costs in their financial records. However, on the basis of the contractor's estimate of contract time spent on 20 GT&E evaluations in fiscal years 2012 through 2015 (including the four evaluations in our sample), we estimated the total cost of these evaluations at approximately \$1.1 million—an average of approximately \$56,300 per evaluation. According to DOD officials, actual costs likely varied across evaluations due to differences in the size of the evaluation teams, the foreign country in which the evaluation took place, and the amount of time each team spent abroad.

- HHS's PEPFAR programs typically conduct evaluations as part of larger cooperative agreements that are not structured to specify evaluation costs. We reviewed cost information for 10 HHS evaluations.²⁹ We identified a specific cost—\$15,400—for only one evaluation, which was conducted under a cooperative agreement specifically for the evaluation; the remaining nine evaluations were conducted as part of cooperative agreements that did not specify evaluation costs. Using budget documents and informed estimates from CDC staff and implementing partners, we estimated that the costs of these nine evaluations ranged from \$25,100 to \$356,200. Additionally, 11 of the 34 HHS evaluations in our sample had no external costs because they were conducted solely by HHS staff using existing datasets. CDC officials stated that CDC intends to track evaluation costs in the future. For example, according to HHS, upon continuation of a cooperative agreement, an implementing partner will be required to report on progress on its Evaluation and Performance Monitoring Plan, as well as on expenditures to date and plans and budgets for the following year.

Evaluation Methods, Period of Performance, and Location Influence Evaluation Costs

Our analysis found that data collection methods, frequency of data collection, evaluation duration, and evaluation location all affect evaluations' cost. For example, evaluations that collected data by surveying program beneficiaries had a median cost of \$202,500—approximately \$74,000 higher than the median cost of those that did not—and evaluations that collected data repeatedly over time had a

²⁹HHS provided GAO with summary evaluation cost estimates based on financial records, budget documents, and informed estimates. Preparing the information required HHS to review a large volume of documents; thus, we limited our review of source documentation to 10 out of the 34 evaluations in our cost sample. See app. I for further information on how we sampled and determined HHS evaluation costs.

median cost of \$194,500—approximately \$44,500 higher than those that did not. Evaluations that took longer to perform also tended to be more expensive. Other factors that might influence costs include unstable locations and evaluations conducted at multiple sites. For example, a performance evaluation conducted in an unstable country cost \$365,700 for 78 days of work, and a performance evaluation that conducted data collection in 12 countries cost \$902,100. In addition, conducting an evaluation in multiple sites within the same country might increase evaluation costs. For example, a performance evaluation conducted in eight cities and seven states in India cost \$407,500, including the cost of a midterm evaluation that was also conducted in multiple cities. These costs greatly exceeded the median costs for all evaluations.

High-Quality Evaluations Tend to Cost More, but Some Lower-Cost Evaluations also Met All Quality Criteria

Our analysis found that high-quality evaluations tend to be more expensive, but well-designed lower-cost evaluations also met the criteria we identified for a high-quality evaluation. Overall, as table 4 shows, the median cost of high-quality evaluations (i.e., evaluations that met all quality criteria) was \$137,800 more than the median cost of acceptable-quality evaluations (i.e., evaluations that partially or generally met all quality criteria) and \$208,600 more than the median cost of lower-quality evaluations (i.e., evaluations that did not meet, or provided insufficient information for, one or more quality criteria).

Table 4: Fiscal Year 2015 MCC, State, USAID, and USDA Foreign Assistance Evaluation Costs, by Quality Category

Dollars in thousands

	Number of evaluations	Cost (dollars)			
		Average	Median	Minimum	Maximum
High quality	15	\$543.3	307.4	97.1	2,230.4
Acceptable quality	39	\$198.8	169.6	9.6	902.1
Lower quality	22	\$159.6	98.8	25.9	981.8

Legend: MCC = Millennium Challenge Corporation, State = Department of State, USAID = U.S. Agency for International Development, USDA = U.S. Department of Agriculture.

Sources: GAO analysis of foreign assistance evaluation reports, evaluation contracts, invoices, and related documents as well as data from Federal Procurement Data System-Next Generation. | GAO-17-316

Note: High-quality evaluations generally met all applicable quality criteria. Acceptable-quality evaluations partially or generally met all applicable quality criteria. Lower-quality evaluations did not meet, or provided insufficient information related to, one or more quality criteria.

High-quality evaluations also tended to include factors associated with higher evaluation costs. For example, the most expensive evaluation in our sample cost \$2.2 million and generally met all quality criteria. This net impact evaluation assessed multiple civil society and governance

programs in an African country using different methodologies, including a randomized controlled trial, over 4 years and conducted two rounds of surveys of program beneficiaries. Another high-quality evaluation cost \$1.4 million and took almost 4 years to complete; this evaluation conducted three surveys of program beneficiaries and used a quasi-experimental methodology to assess the net impacts of energy-efficient stoves in Asia. However, some lower-cost evaluations also met all of the quality criteria. Of the 15 high-quality evaluations for which we identified costs, 4 cost less than \$150,000.

Selected Agencies' Evaluations Are Generally Available Online, but Some Agencies Can Improve Dissemination

We assessed DOD's, HHS's, MCC's, State's, USAID's, and USDA's use of six dissemination practices that federal, AEA, and other guidance indicate agencies should generally use to ensure effective dissemination of evaluations.³⁰ We found that the agencies varied in their performance of these practices for the fiscal year 2015 evaluations we reviewed (see table 5). All except USDA generally made nonsensitive evaluations publicly available online.³¹ These nonsensitive evaluations could generally be located with the agencies' website search engines. However, some agencies' evaluations were not posted in a timely manner. Each of the agencies posted its sensitive evaluations internally for access by internal users. Only USAID included dissemination plans in most nonsensitive evaluations to help ensure their dissemination to potential users of the evaluation, but most of the other agencies now require such plans to be prepared for future evaluations. In addition to publicly posting the report, all of the agencies used other means to actively disseminate evaluation findings. Following these practices can help agencies ensure that their evaluation reports are accessible, timely, and useful to decision makers and other stakeholders.

³⁰We identified these practices through a review of federal, AEA, and other guidance. For example, federal guidance requires that agencies report their performance publicly in an open, transparent, evidence-based method; see U.S. Office of Management and Budget (OMB), *Preparation, Submission, and Execution of the Budget*, Circular No. A-11 (Washington, D.C.: June 2015). Presidential Policy Directive-6 on U.S. Global Development Policy directs agencies to incorporate evaluation findings in policy and budget decisions; see *Presidential Policy Directive 6* (Washington, D.C.: Sept. 22, 2010). The AEA Roadmap encourages evaluators to disseminate the findings of evaluation reports to relevant stakeholders; see American Evaluation Association, *An Evaluation Roadmap for a More Effective Government* (October 2013).

³¹Sensitive evaluations are deemed to contain information with the potential to cause foreseeable harm to governmental, commercial, or private interests if disseminated to the public or persons who do not need the information to perform their jobs. Nonsensitive evaluations are those that an agency does not designate as containing such information.

Table 5: Assessment of Six U.S. Agencies' Use of Six Practices for Effective Dissemination of Foreign Aid Evaluations for Fiscal Year 2015

Effective dissemination practices ^a	DOD	HHS	MCC	State	USAID	USDA
1. Generally post nonsensitive evaluations online	N/A ^b	Yes	Yes	Yes	Yes	No ^d
2. Provide a search engine that can find the evaluations	N/A	Yes ^c	Yes	Yes	Yes	N/A ^d
3. Post evaluations in the timeframe required by the agency	N/A	No	No	N/A ^e	Yes	N/A
4. Make sensitive evaluations accessible internally	Yes	N/A ^f	N/A ^f	Yes	Yes	N/A ^f
5. Require planning for the dissemination of evaluations	Yes ^g	Yes	Yes	No	Yes	Yes
6. Use means other than public posting to disseminate evaluations	Yes ^g	Yes	Yes	Yes	Yes	Yes

Legend: DOD = Department of Defense, HHS = Department of Health and Human Services, MCC = Millennium Challenge Corporation, State = Department of State, USAID = U.S. Agency for International Development, USDA = U.S. Department of Agriculture, N/A = not applicable.

Source: GAO analysis of agency evaluation guidance, evaluation and dissemination documents, and websites. | GAO-17-316

^aWe identified six dissemination practices, based on federal and other guidance, which agencies should generally use to ensure the effective dissemination of evaluation reports.

^bThe four DOD evaluations we reviewed, for the Global Train and Equip (GT&E) program, were designated "sensitive" and thus were not required to be posted on a public website.

^cThe Centers for Disease Control and Prevention (CDC) is in the process of adding all evaluations from fiscal year 2015 to the CDC Stacks website. This website has a search engine that can be used to locate individual evaluations. CDC reported that 49 of the 51 evaluations from fiscal year 2015 had been posted on the CDC Stacks website as of December 2016.

^dUSDA does not post evaluations online. According to USDA officials, the department is in the process of developing procedures for making nonsensitive evaluations public.

^eIn fiscal year 2015, State did not have a policy requiring that evaluations be posted within a certain timeframe. In fiscal year 2016, State revised its guidance to require that evaluations be posted online 90 days after completion.

^fHHS, MCC, and USDA did not designate as sensitive any of the evaluations we reviewed.

^gDOD plans for evaluation dissemination by identifying potential users of the evaluation on a standard e-mail distribution list. DOD uses this standard e-mail distribution list to disseminate GT&E program evaluations via e-mail to congressional stakeholders, as required by law, as well as internal stakeholders.

**All Agencies except USDA
Make Nonsensitive
Evaluations Publicly
Available Online**

Every agency with nonsensitive evaluations requires public, online posting of nonsensitive evaluation documents, and all except USDA publicly posted all of the nonsensitive evaluations we reviewed on publicly accessible websites. Making evaluation reports publically available on their websites helps agencies share evaluation findings with partners, program beneficiaries, and the wider public and facilitates the incorporation of evaluation findings into program management decisions.

Agency Websites for Foreign Aid Evaluations

Among the nonsensitive evaluations we reviewed, those conducted by the Department of Health and Human Services (HHS) for the President's Emergency Program for AIDS Relief (PEPFAR); the Millennium Challenge Corporation (MCC); the Department of State (State); and the U.S. Agency for International Development (USAID)—were posted on publicly accessible websites.

HHS PEPFAR

<https://data.pepfar.net/evaluations> and <https://stacks.cdc.gov/>

MCC

<https://data.mcc.gov/evaluations/index.php/catabg>

State

<http://www.state.gov/evaluations/all/index.htm>

USAID

<https://dec.usaid.gov>

Source: GAO analysis of agency websites. | GAO-17-316

We examined the agencies' dissemination of 193 evaluations. The agencies did not require 22 evaluations to be publicly posted due to their sensitivity—all 4 DOD evaluations as well as 17 evaluations from State and 1 from USAID. Of the remaining 171 nonsensitive evaluations, we found that more than three-quarters (133) were publicly posted.³² USDA did not publicly post any of its 38 nonsensitive evaluations. According to USDA officials, the department is in the process of developing procedures for making these nonsensitive evaluations public, which would include reviewing the documents to ensure that they did not contain, for example, personally identifiable or proprietary information. Without posting all nonsensitive evaluations online, agencies cannot ensure that the evaluations' findings reach intended audiences and are available to inform future program design or budget decisions.

Publicly Posted Evaluations Can Generally Be Found with Agency Search Engines

Most of the nonsensitive, publicly posted evaluations we reviewed could be located with a search engine on the agencies' websites. Providing a search engine that potential evaluation users can employ to find the evaluation reports ensures that users can locate the information they seek, in a format that matches their expectations.³³ Websites at three of the four agencies with publicly posted evaluations—MCC, State, and USAID—have search engines that enable users to find each specific evaluation.³⁴ The PEPFAR website, which hosts evaluations of PEPFAR

³²A small number of the evaluations we reviewed lacked key information that would enable a user to assess the strength of the evaluation's evidence. Two of the 63 USAID evaluations we reviewed were posted without cited appendixes describing the methods used in the evaluation, which would allow a reader to assess the validity of the findings. Likewise, 3 of the 38 evaluations we obtained from USDA were missing appendixes that the evaluations cited as providing information about the methods used. State provided only a partially illegible scanned copy of one of its sensitive evaluations from fiscal year 2015.

³³Department of Health and Human Services, *Research-Based Web Design and Usability Guidelines* (Washington, D.C.: 2006), accessed November 30, 2016, https://www.usability.gov/sites/default/files/documents/guidelines_book.pdf.

³⁴The MCC, CDC, and USAID websites also allow users to search for evaluations using additional filter criteria, such as the country associated with the evaluation or the year in which the evaluation was completed. State's website does not offer additional filter criteria, but State posted relatively few (six) publicly available evaluation reports in fiscal year 2015.

programs implemented by CDC and other agencies, has these evaluations listed in a spreadsheet locatable on the site. Many evaluations of PEPFAR programs implemented by CDC can also be found using a search engine at a separate website, called “CDC Stacks.” According to CDC, almost all of the evaluations that we reviewed were posted on the CDC Stacks website. The agency reported that it is in the process of adding the remaining CDC evaluations from fiscal year 2015 to this website.

Some Evaluations Were Not Posted within Required Time Frames

Some of the nonsensitive evaluations we reviewed were not posted on the agencies’ websites within required timeframes. Making evaluation reports accessible in a timely manner ensures that interested parties can access the findings of these evaluations in time to incorporate them into program management decisions. MCC and HHS did not post some evaluations within the timeframes they require, limiting stakeholders’ ability to make optimal use of the evaluation findings.³⁵

We found that MCC did not post 10 of its 16 evaluations, as MCC requires, within 6 months after MCC received them, and it did not post 8 of these 10 evaluations until a year or more after MCC received them. According to MCC officials, the agency’s internal evaluation quality review process for evaluations, in which the agency reviews the document before releasing it to the public, has been a major factor in these delays. MCC officials reported that for some of the evaluations—for instance, those written in a language other than English—this process took significantly longer than usual.

HHS did not post 11 HHS evaluations in the timeframe required by the agency. It did not post 6 of these 11 evaluations online within 90 days as required by PEPFAR. PEPFAR guidance requires that evaluations be posted within 90 days of completion, while HHS requires that evaluations be publicly posted within a year of their completion. One HHS official stated that the delay in the posting of these six evaluations was due to the conflicting policies. However, the remaining five evaluations were also not

³⁵We have previously recommended that DOD take steps to develop a process for improving the timely completion and submission of required GT&E evaluation reports to Congress. DOD submitted its fiscal year 2012 GT&E evaluation report to Congress in accordance with required deadlines. However, DOD submitted its fiscal year 2013 report to Congress in September 2015, 21 months later than required, and submitted its fiscal year 2014 report to Congress in December 2015, 12 months later than required. DOD’s report for fiscal year 2015 was 1 month late. See [GAO-16-368](#).

posted within the year as required by HHS/CDC. These five evaluations have since been posted online. Since evaluated conditions may change over time, not posting evaluations online within the required timeframe limits internal and external stakeholders' access to current, actionable information. In comments on a draft of this report, CDC noted that, as of December 2016, CDC is providing guidance that all evaluations be posted online within 90 days, as required by PEPFAR. CDC published this guidance in January 2017.

All Agencies Have Websites to Make Sensitive Evaluations Available Internally

Of the three agencies with sensitive evaluations—DOD, State, and USAID—all have websites to make these evaluations available to internal stakeholders. While sensitive evaluations are not required to be made available to the public on an agency's website, disseminating sensitive evaluation findings to the appropriate audience will facilitate their use. DOD, State, and USAID all reported that they have internal websites that can be used to post sensitive evaluations. In addition, State updated its policy for 2015 to require that State officials post a nonsensitive summary of sensitive evaluations on State's public website. While USDA does not currently publicly post its evaluations, USDA reported that it makes these evaluations internally available through its grant management system. HHS and MCC did not have sensitive evaluations in fiscal year 2015.

USAID Included Dissemination Plans in Most Evaluations, and Other Agencies Will Require Such Plans in the Future

USAID requires the development of dissemination plans and included evidence of such planning in the majority of the evaluations we reviewed, and all of the other agencies except State now require such plans for nonsensitive evaluations. Dissemination planning identifies potential users of an evaluation and describes an approach to providing users with the evaluation results.³⁶ Such planning can help agencies ensure that evaluation reports are disseminated effectively

Among the six agencies, only USAID required the development of dissemination plans for fiscal year 2015 evaluations and included evidence of such planning in the majority of the evaluations whose dissemination we reviewed. Of the 62 USAID evaluations, 44 included evidence that dissemination planning had been completed. HHS, MCC, State, and USDA did not require dissemination plans for their evaluations

³⁶Organization for Economic Co-operation and Development (OECD), *Evaluating Development Activities: 12 Lessons from the OECD DAC* [Development Assistance Committee] (Paris, France: 2013).

completed in fiscal year 2015. Agency officials at HHS and MCC provided evidence that dissemination planning took place for at least one of their respective evaluations we reviewed, but this dissemination planning was not required by agency policy, and therefore this planning was ad hoc. DOD plans for evaluation dissemination by identifying potential users of the evaluation and sending e-mails to these internal and congressional stakeholders after the evaluations are completed.

HHS, MCC, and USDA guidance now require dissemination plans for future evaluations. State officials reported that, as of November 2016, State was planning to revise its policy to require the use of dissemination plans for evaluations but had not instituted this requirement. Without dissemination planning, State cannot ensure that its evaluations are disseminated as effectively as possible to potential users.

Agencies Used Additional Means to Actively Disseminate Evaluation Findings

In addition to posting evaluations online, each of the six agencies reported disseminating evaluation findings through other means. Our prior work has shown that taking such additional steps to actively disseminate evaluation reports—for example, briefing stakeholders on evaluation findings and distributing the evaluations to interested stakeholders via e-mail—facilitates dissemination of evaluation report findings and encourages their use.³⁷

Agency officials reported using various means besides web posting to disseminate evaluation findings. For example, officials at all six agencies reported using briefings to share evaluation findings with various stakeholders within and outside of the agency. Additionally, HHS, State, MCC, and USAID officials reported that they shared evaluation results with interested parties at various professional conferences. USAID officials stated that the agency also disseminates evaluation findings by posting its evaluations on partner websites, creating video companions to evaluation reports to provide to stakeholders, and posting syntheses of evaluation findings on the agency's website.

³⁷GAO, *Program Evaluation: Strategies to Facilitate Agencies' Use of Evaluation in Program Management and Policy Making*, [GAO-13-570](#) (Washington, D.C.: June 26, 2013).

Conclusions

Foreign assistance evaluations can be challenging to implement, but they are an essential tool for guiding agency decision making and allocation of resources. Agencies' foreign assistance evaluations assess a wide variety of programs around the world, using many different designs and methodologies, and the wide range of evaluation costs reflects this diverse context. However, regardless of the location, design, or cost, an evaluation should provide sufficient and reliable evidence to support its findings. A high-quality evaluation helps agencies and stakeholders identify successful programs to expand or pitfalls to avoid. Evaluations that do not meet all quality criteria that we identified may not provide sufficiently reliable evidence to inform these decisions. In addition, for evaluations to inform decision making, stakeholders must be able to find them. While foreign assistance agencies have generally made their evaluations available online in a timely manner, several agencies can take additional steps to ensure that stakeholders have improved access to these evaluations to make better-informed decisions about future program design and implementation. A growing body of high-quality, broadly disseminated evaluations can help the United States continuously improve its foreign assistance programs and thereby support democracy, enhance security, reduce poverty and suffering, and achieve other U.S. foreign policy goals.

Recommendations for Executive Action

To improve the reliability and usefulness of program evaluations for agency program and budget decisions, we recommend that the Chief Executive Officer of MCC, the Administrator of USAID, the Secretary of Agriculture, the Secretary of Defense, the Secretary of State, and the Secretary of Health and Human Services (in cooperation with State's Office of the U.S. Global AIDS Coordinator and Health Diplomacy) each develop a plan for improving the quality of evaluations for the programs included in our review, focusing on areas where our analysis has shown the largest areas for potential improvement.

To better ensure that the evaluation findings reach their intended audiences and are available to facilitate incorporating lessons learned into future program design or budget decisions, we recommend that

- the Secretary of Health and Human Services direct the Centers for Disease Control and Prevention to update its guidance and practices on the posting of evaluations to require PEPFAR evaluations to be posted within the timeframe required by PEPFAR guidance;

-
- the Chief Executive Officer of MCC adjust MCC evaluation practices to make evaluation reports available within the timeframe required by MCC guidance;
 - the Secretary of State amend State's evaluation policy to require the completion of dissemination plans for all agency evaluations; and
 - the Secretary of Agriculture implement guidance and procedures for making FAS evaluations available online and searchable on a single website that can be accessed by the general public.

Agency Comments and Our Evaluation

We provided a draft of this report to DOD, State, HHS, MCC, USAID and USDA for review and comment. DOD, State, HHS, MCC, USAID, and USDA provided official comments, which are reproduced in appendixes III through VIII with, where relevant, our responses. DOD, HHS, and USAID also provided technical comments, which we incorporated as appropriate.

The following summarizes DOD, State's, HHS's, MCC's, USAID's, and USDA's official comments and our responses.

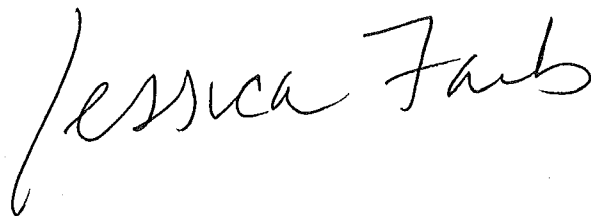
- DOD stated that it partially concurred with our recommendation, noting that in many cases, certain methodologies are not well suited for security assistance evaluation. DOD observed that, for example, it would be unethical to establish randomized control groups for security assistance evaluations and that foreign military organizations may be unwilling to provide DOD significant access to some military units solely for the purpose of the evaluation. We recognize that certain methodologies are not appropriate in every context, and we did not advocate the use of randomized control groups in the DOD evaluations we reviewed. Our main concerns about the DOD evaluations focused on implementation of the methods used. In particular, we found limitations in sampling methods including descriptions of the target population, data collection methods, and data analysis. We adjusted pertinent wording in our report to clarify these points.
- State concurred with our recommendations and noted that its forthcoming Program Design and Performance Management Policy for Programs, Projects, and Processes, recently published Program Design and Performance Management toolkit, and updated policy guidance will constitute a plan for improvement. We will monitor the implementation of this plan to verify that State takes appropriate steps to address our recommendation.

-
- HHS concurred with our recommendation that it update guidance and practices on the posting of PEPFAR evaluations and stated that CDC guidance now requires evaluation reports to be posted on a publically accessible website within 90 days of the evaluation's completion. HHS did not comment on our recommendation that it develop a plan for improving the quality of evaluations.
 - MCC stated that it welcomed our findings and recommendations for improvement but noted that it could not agree or disagree with our quality assessments because we did not provide data on our determinations for individual evaluations. In response to our observation that MCC evaluations did not contain conflict-of-interest statements, MCC noted that it has required independent third-party evaluation of all its projects since 2009 and that, in 2013, it standardized the language in its evaluation contracts to explicitly establish the independent role of evaluators. While these are positive steps, we believe that including in MCC's published evaluations explicit statements about the evaluators' independence and any potential conflicts of interest would bolster the evaluations' credibility and usefulness. With regard to the timeliness of public access to its evaluations, MCC indicated that when it established its internal review process for evaluations in 2013, it did not anticipate the length of time that would be required to finalize evaluation reports. MCC noted that its forthcoming revised policy on monitoring and evaluation states that "MCC expects to make each interim and final evaluation report publicly available as soon as practical after receiving the draft report." However, the revised policy does not establish a target time frame for completing internal reviews of the reports. Establishing such a time frame could help MCC ensure that evaluation reports are published in a timely fashion that maximizes their usefulness.
 - USAID stated that it has established a plan to improve the quality of evaluations, including an update and clarification of the requirements and quality standards for evaluations. USAID also stated that it plans to provide additional training and other capacity-building efforts to help ensure that staff have the necessary skills to manage evaluations. We will monitor implementation of this plan to verify that USAID takes appropriate steps to address our recommendation.
 - USDA agreed with our recommendations. To address the recommendations, USDA stated that it would update its guidance on reviewing evaluation terms of reference to include a section on quality that specifically focuses on the areas where the GAO analysis has

shown the largest areas for potential improvement. USDA further stated that FAS will continue its current efforts to make nonsensitive evaluations publicly available online and will make them searchable as well.

We are sending copies of this report to the appropriate congressional committees and to the Secretaries of Agriculture, Defense, State, and Health and Human Services; the Chief Executive Officer of the Millennium Challenge Corporation; and the Administrator of the U.S. Agency for International Development. In addition, the report will be available at no charge on GAO's website at <http://www.gao.gov>.

If you or your staff have questions about this report, please contact me at (202) 512-6991, or farbj@gao.gov. Contact points for our Offices of Congressional Relations and Public Affairs may be found on the last page of this report. GAO staff who made major contributions to this report are listed in appendix IX.

A handwritten signature in black ink that reads "Jessica Farb". The signature is written in a cursive, flowing style.

Jessica Farb
Acting Director, International Affairs and Trade

List of Requesters

The Honorable Bob Corker
Chairman
Committee on Foreign Relations
United States Senate

The Honorable Ed Royce
Chairman
The Honorable Eliot Engel
Ranking Member
Committee on Foreign Affairs
House of Representatives

The Honorable Ted Poe
Chairman
Subcommittee on Terrorism, Non-Proliferation, and Trade
Committee on Foreign Affairs
House of Representatives

The Honorable Gerald Connolly
Ranking Member
Subcommittee on Government Operations
Committee on Government Oversight and Reform
House of Representatives

The Honorable Ted Cruz
United States Senate

Appendix I: Objectives, Scope, and Methodology

In response to congressional requests, we examined (1) the extent to which foreign assistance program evaluations met key evaluation quality criteria; (2) the costs of the evaluations, as well as factors that affect these costs; and (3) the extent to which the agencies ensure the dissemination of evaluation reports within the agency and to the public.

To address our objectives, we identified the six major agencies administering the most foreign assistance on the basis of obligations reported to the U.S. Agency for International Development's (USAID) U.S. Overseas Loans and Grants database for fiscal years 2008 through 2012. The six agencies we identified are USAID, the Department of State (State), the Millennium Challenge Corporation (MCC), the Department of Health and Human Services (HHS), the U.S. Department of Agriculture (USDA) and the Department of Defense (DOD). For the three agencies that are not focused exclusively on foreign aid or foreign affairs (HHS, USDA, and DOD), we limited our scope to selected programs. For HHS and USDA, we selected programs that account for the vast majority of foreign assistance program dollars that the agency implemented. At HHS we examined evaluations of the President's Emergency Plan for AIDS Relief (PEPFAR) programs implemented by HHS's Centers for Disease Control and Prevention (CDC). At USDA we examined evaluations for the Food for Progress and McGovern-Dole food assistance programs, implemented by the Foreign Agricultural Service (FAS). At DOD we examined evaluations prepared for the Global Train and Equip (GT&E) program. While our previous review of agency evaluation policies did not identify DOD-wide evaluation policies,¹ we did identify GT&E as having relevant policies to guide its evaluations.

To identify evaluations completed in fiscal year 2015, the most recently completed fiscal year as we undertook our review, we requested that each agency provide a list of all foreign aid evaluation reports completed in that year. We did not separately review agency files to identify if agencies had additional evaluations beyond those listed in the registries. To assess the reliability of the agency evaluation lists, we reviewed the documents provided to ensure that each was a completed evaluation and to confirm that the date of the document fell within our specified

¹GAO, *Foreign Assistance: Selected Agencies' Monitoring and Evaluation Policies Generally Address Leading Practices*, [GAO-16-861R](#) (Washington, D.C.: Sept. 27, 2016). DOD issued an agency-wide policy in January 2017. See DOD Instruction 5132.14, *Assessment, Monitoring, and Evaluation Policy for the Security Cooperation Enterprise* (Jan. 13, 2017), available at <http://www.dtic.mil/whs/directives>.

timeframe. If necessary, we followed up with agency officials to clarify the date or status of the document. Based on their responses, we removed documents that were not evaluations or fell outside of our timeframe. We also did not review evaluation reports that were not written in English. We determined that the data in the evaluation lists were sufficiently reliable for the purposes of this engagement. In all, we identified a study population of 361 evaluations: 4 DOD evaluations, 51 HHS evaluations, 17 MCC evaluations, 28 State evaluations, 221 USAID evaluations, and 40 USDA evaluations. We examined the evaluations themselves and any appendices that the agency provided which were directly referred to in the evaluations. We did not consider evaluation plans and protocols, underlying documents and other work papers as evidence that the planned design was implemented. Similarly, we did not consider contracts with third-party evaluators or evaluation organizations as evidence that the evaluator had maintained independence. Instead we required statements in the reports or methodological appendices that steps and procedures were actually taken and that no threats to independence have been identified.

From the study population of fiscal year 2015 evaluations, we reviewed all DOD, MCC, and State evaluations; all USAID net impact evaluations; and a sample of HHS, USDA, and USAID performance evaluations. We randomly selected a probability sample from the study population of HHS, USDA, and USAID performance evaluations. With this probability sample, each member of the study population had a nonzero probability of being included, and that probability could be computed for any member. For USAID, we included all net impact evaluations in the sample because net impact evaluations constituted less than 20 percent of all the evaluations provided, and if we had not included them all, we would not have been able to comment on this type of evaluation.²

Based on the review of the evaluation documents after the initial screening, an additional 16 evaluations were found not to be within our scope, and we substituted for these evaluations when possible. For example, we excluded documents that did not evaluate a specific program, were monitoring or grant reports, or were plans for an evaluation rather than an evaluation report. We included two substitute HHS and two substitute USDA evaluations to replace those that were

²The evaluations we reviewed for State and MCC also included net impact evaluations. HHS, DOD, and FAS did not categorize any of their fiscal year 2015 evaluations as net impact evaluations.

found to not be in scope and also reviewed additional USDA evaluations. However, because we had initially included all MCC, State, and USAID net impact evaluations, there were no additional evaluations available to substitute if those were excluded. The original sample and the final respondents across the six agencies can be found in table 6. Each sample selection was subsequently weighted in the analysis to represent the evaluations in the population that were not selected.

Table 6: Agency Foreign Assistance Evaluation Study Population and Sampling

Agency	Study population	Planned sample	Final sample
DOD	4	4	4
HHS	51	34	34
MCC	17	17	16
State	28	28	23
USAID	201 performance	49 performance	49 performance
	20 net impact	20 net impact	14 net impact
USDA	40	28	33
Total	361	180	173

Legend: DOD = Department of Defense, HHS = Department of Health and Human Services, MCC = Millennium Challenge Corporation, State = Department of State, USAID = U.S. Agency for International Development, USDA = U.S. Department of Agriculture.

Source: GAO analysis of fiscal year 2015 agency evaluation data. | GAO-17-316

We reviewed the full population of DOD, MCC, and State evaluations; therefore, our results from the quality review of these evaluations do not have an associated margin of error. The results from our review of the HHS, USDA, and USAID evaluations are reported with an associated margin of error. Because we followed a probability procedure based on random selections, our sample is only one of a large number of samples that we might have drawn. Since each sample could have provided different estimates, we express our confidence in the precision of our particular sample's results as a 95-percent confidence interval. This is the interval that would contain the actual population value for 95 percent of the samples we could have drawn. All percentage estimates for aggregated results from our review have margins of error at the 95 percent confidence level of plus or minus 8 percentage points or less, unless otherwise noted, and all percentage estimates for individual agencies from our review have margins of error at the 95 percent confidence level of plus or minus 11 percentage points or less, unless otherwise noted.

To assess the extent to which the results of foreign assistance program evaluations are supported by their evidence and whether they assess if programs have met their goals, we assessed the sample of agency fiscal year 2015 evaluation reports against quality criteria we identified. We identified these criteria based on our review and analysis of evaluation guidance from agencies included in our review (including any agency internal evaluation review checklists), international organizations,³ evaluation organizations,⁴ and prior GAO reporting.⁵ These criteria include necessary high-level elements in designing, implementing and reporting on evaluations that could serve as standards across different agencies and evaluation types. Prior to undertaking our quality review, these criteria were discussed and reviewed within the engagement team, as well as by other GAO staff with experience in program evaluation and methodologies.

We incorporated the identified criteria into a standardized data collection instrument (DCI) in order to consistently review the sampled evaluation reports. The DCI contained evaluative questions against which to assess evaluation quality as well as descriptive questions to gather information about the evaluations, such as its location and methodology. The high-level criteria each included subquestions about elements the reviewer should consider in making his or her overall decision. The evaluation quality criteria were judged on a four-part scale for most of the judgmental questions:

³See, for example, United Kingdom Department for International Development, *International Development Evaluation Policy* (May 2013), https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/204119/DFID-Evaluation-Policy-2013.pdf, and Organization for Economic Co-operation and Development (OECD), *Development Assistance Committee Guidelines and Reference Series, Quality Standards for Development Evaluation* (Paris, France: April 2010), available at <http://www.oecd.org/dataoecd/55/0/44798177.pdf>.

⁴For example, American Evaluation Association (AEA), *An Evaluation Roadmap for a More Effective Government*, accessed November 10, 2016, <http://www.eval.org/d/do/472>.

⁵See GAO, *President's Emergency Plan for AIDS Relief: Agencies Can Enhance Evaluation Quality, Planning, and Dissemination*, [GAO-12-673](#) (Washington, D.C.: May 31, 2012); GAO, *Designing Evaluations: 2012 Revision*, [GAO-12-208G](#) (Washington, D.C.: January 2012); GAO, *Water and Sanitation Assistance: USAID Has Increased Strategic Focus but Should Improve Monitoring*, [GAO-16-81](#) (Washington, D.C.: Oct. 6, 2015); and GAO, *International Cash-Based Food Assistance: USAID Has Established Processes to Monitor Cash and Voucher Projects, but Data Limitations Impede Evaluation*, [GAO-16-819](#) (Washington, D.C.: Sept. 20, 2016).

- *generally addressed*: the evaluation mostly addressed the key element(s) of the criterion but did not have to completely address all elements in the subquestions;
- *partially addressed*: the evaluation had one or more clear area(s) for improvement on the criterion;
- *not at all addressed*: the evaluation did not show that steps were taken to address the criterion; and
- *insufficient information*: reviewers could not make a determination due to a lack of information in the evaluation and any other associated materials.

If a criterion was partially or not at all addressed, or if there was insufficient information in the evaluation to assess the criterion, we considered it a deficiency. We did not consider study protocols or design documents that indicated plans for a particular evaluation step as sufficient evidence that such a step was performed unless the evaluation report also provided evidence that it had.

The descriptive questions in the DCI about evaluation types and methodology were based on prior GAO work⁶ and asked about designs that examined net impacts of interventions, outcomes of interventions, and processes. The questions on net impact evaluations asked about the type of design using four categories: (1) randomized controlled trials or groups, (2) comparison groups, (3) time series that would allow for trends to be determined pre- and post- intervention, and (4) quasi-experimental statistical modelling techniques. The questions on outcome and process evaluations asked whether baselines and targets had been established for the outcomes and whether criteria had been established to assess processes. From these types of evaluation, we created two broad categories to use in our analysis of evaluation cost and quality: net impact evaluations, and performance evaluations. The net impact category included all four impact design types, while the performance evaluation category included outcome and process evaluations as well as a few evaluations that did not fall within the outcome and process categories. The performance evaluations included some that had established targets or baselines and others that had not, while the process evaluations included some that had established criteria for assessment and others that had not. We noted some overlap between the net impact,

⁶See [GAO-12-208G](#) and GAO, *Performance Measurement and Evaluation: Definitions and Relationships*, [GAO-11-646SP](#) (Washington, D.C.: May 2011)

performance, and process categories. For example, net impact evaluations often considered outcomes or processes, and performance evaluations often considered both outcomes and processes. This overlap was a key reason we decided to develop and analyze two broad types of evaluation categories rather than attempt to develop more refined types.

A key assumption underlying our analysis was that different types of evaluations were appropriate for different types of study questions. The evaluative questions in the DCI were relative, rather than absolute, with respect to the study questions and were intended to be applicable across evaluation types. We did not assess the study questions in terms of their scope or rigor. We instead took the study questions as given, thereby giving every evaluation an equal chance to receive a high score if its design and implementation were appropriate for the study questions. Evaluation reviewers were instructed to consider design, implementation, and reporting in terms of the study questions the evaluations set out to answer rather than against an absolute standard. In this way, we determined that it would be as possible for a qualitative midterm evaluation that considered program implementation to achieve high scores as it would for a final net impact evaluation that considered effects attributable to the program.

The main criteria questions in the DCI asked about the appropriateness of the design, implementation, and conclusions in light of the study objectives. We did not determine a single definition of appropriateness because we recognized that it is dependent on the study objectives and data collection conditions. For example, the standards for appropriateness of a final net impact evaluation of a pilot health care program that seeks to establish whether the intervention is achieving positive outcomes are different from the standards for a mid-term performance evaluation of a well-established water and sanitation program supported by a solid evidence base and focused on whether the program was implemented as planned. Given the variation in agency goals and programs, evaluation types, and evaluation timing, we determined that we would rely on expert professional judgment rather than attempt to use a single definition of appropriateness for every situation.

While we designed our DCI to apply broadly across agencies and evaluation types, differences in agency evaluation practices and areas of responsibility may limit comparisons between the agencies. For example, the target audience of an evaluation may determine whether it includes certain reporting elements. HHS's PEPFAR evaluations are generally

produced for dissemination in research publications and journals, while USAID evaluations evaluate a wide range of programs and are generally directed to an audience of program officials and managers. In addition, the evaluations we reviewed assessed a wide range of foreign assistance programs with varying characteristics. These characteristics include the nature of the foreign assistance intervention, the type of program responsible for the intervention, whether the program was designed to be evaluated, and the timing of the evaluation. For example, some evaluations consider ongoing development assistance in areas such as education or health, while others consider emergency responses to humanitarian crises. Agency officials noted that it could be harder to ensure quality in an evaluation of a program that had to respond quickly to a crisis and therefore did not have the opportunity to plan for an evaluation. They also noted that if an evaluation is not started until after the program has begun, there may not be any baseline data available.

The evaluation review consisted of multiple reviews by a team of GAO staff with experience and familiarity with research methods as well as with reviewing studies and evaluations across a wide range of subject areas and disciplines. After completing his or her initial review, the first reviewer notified the second reviewer that the evaluation report was available for his or her review. The second reviews were not independent; as the second reviewer saw the decisions made by the first reviewer and could review the first reviewer's notes on sources and justifications for his or her decisions. The second reviewers read the evaluation and indicated whether he or she agreed with the first reviewers' decisions or whether he or she proposed another decision. The first and second reviewers subsequently met to reconcile any differences. After the reconciliations were completed, a supervisor then reviewed the work of the two reviewers for internal consistency and completeness according to a standard protocol but did not re-review the evaluation documents. The supervisor related the identified issues as needed to the first and second reviewers, who addressed them before the supervisor recorded the review as final.

We took several steps to ensure consistency among the reviewers. We conducted two pretests of the DCI on sample evaluations. The first pretest included members of the engagement team as well as GAO staff with experience in the design of survey instruments or in the review of foreign assistance evaluations. The second pretest included members of the evaluation review team. After each round of pretests, we made appropriate revisions to the DCI. To help ensure consistency of interpretation, we created a guidance document where reviewers

recorded questions about certain decision rules to follow in specific instances. Answers to the questions were then posted after discussion among the review and engagement team members. Additionally, the engagement and review team held regular weekly meetings to discuss any methodological issues that arose and preliminary tabulations of the review data.

To analyze the responses to the DCI, we examined how evaluations' quality varied by the eight quality criteria, by agency, by timing of the evaluation relative to the stage of program implementation (midterm or interim vs. final), and type of evaluation (net impact vs. performance). While there were some differences between the final and midterm evaluations, these were not statistically significant. However, a higher percentage of evaluations that attempted to assess net impacts were of high quality than those that did not attempt to assess net impacts. We determined that these differences were due primarily to specific weaknesses in the implementation of the design of the evaluations. For example, performance evaluations used nonrandom sampling more often than net impact evaluations, which used at least some random sampling. While our DCI and subsequent analysis treated both methods of sampling equally, we focused our assessment on the extent to which each method had been appropriate for the study questions. We found that the performance evaluations' nonrandom sampling was carried out appropriately less often than the random sampling typically used in net impact evaluations. In the body of our report, therefore, we focus on the specific weaknesses that we found, such as the one regarding nonrandom sampling, rather than differences at the level of the two broad evaluation types.

Our analysis found a reasonably high degree of overlap between several approaches that we considered for categorizing the evaluations. For example, we categorized the evaluations into three groups: *high quality*, *acceptable quality*, and *lower quality* based on the number of quality criteria that were generally or partially met for each evaluation, as well as instances when quality areas were either not met or there was insufficient information to determine if a certain criterion was met by a particular evaluation. Those evaluations that fell into the lowest and middle categories based on these three categories also generally fell into the lowest and middle categories using another approach that we examined, as table 7 shows. This comparison also shows that some evaluations in the highest category had a relatively higher number of criteria generally met, while some in the middle category had a relatively lower number of criteria generally met.

Table 7: Number of Quality Criteria Generally Met, by Quality Category

Category	Number of quality criteria generally met			Totals
	0-4	5-6	7	
<i>High quality</i> : Generally met all applicable criteria.	0	0	48	48
<i>Acceptable quality but could be improved</i> : Generally or partially met all applicable criteria but did not generally meet all.	25	50	1	76
<i>Lower quality</i> : Did not meet, or provided insufficient information related to, one or more applicable criteria.	35	14	0	49
Totals	60	64	49	173

Source: GAO analysis of fiscal year 2015 agency evaluation data. | GAO-17-316

Note: This analysis groups the evaluations according to the three quality categories we developed based on all eight quality criteria. However, the scores are based on the seven criteria that all 173 evaluations had in common; it does not include the recommendations and lessons learned criterion. One evaluation generally met the seven criteria common to all evaluations but did not generally meet the recommendations and lessons learned criterion.

To determine the cost of foreign aid evaluations, we reviewed contracts, invoices, and related documents to determine the cumulative cost of final evaluations that were conducted by an outside evaluator. We defined the cumulative cost as the cost of the final evaluation and any related activities that informed the evaluation’s findings, such as a separate data collection effort or a midterm or baseline evaluation. We did not determine the cost of evaluations that had a midterm evaluation but not a final evaluation because midterm evaluation costs do not reflect the total cumulative cost of evaluating a program. State and USAID officials noted that some of their midterm evaluations may ultimately be a program’s only evaluation. We did not review the State and USAID fiscal year 2015 midterm evaluations to determine if the agency intends to conduct an additional, final evaluation or include these midterms in our cost analysis. Table 8 shows the total number of evaluations in GAO’s quality sample, the number of those evaluations whose costs we reviewed, and the number of evaluations whose costs we reviewed that we included in the statistical analysis by agency.

Table 8: Number of Evaluations in Cost Sample, by Agency

Agency	Number of evaluations		
	Quality sample	Costs reviewed	Costs included in statistical analysis
DOD	4	4	0
HHS	34	34	0
MCC	16	13	12
State	23	17	16
USAID	63	40	38
USDA	33	10	10
Total	173	118	76

Legend: DOD = Department of Defense; HHS = Department of Health & Human Services; MCC = Millennium Challenge Corporation; State = Department of State; USDA = U.S. Department of Agriculture; USAID = U.S. Agency for International Development.

Source: GAO analysis of foreign assistance evaluation reports, evaluation contracts, invoices, and related documents, as well as data from Federal Procurement Data System-Next Generation. | GAO-17-316

We used contracts and invoices to determine the cost of 42 of the 76 MCC, State, USAID, and USDA final evaluations. To determine an evaluation’s cost, we used either the final invoiced amount or the contract’s total obligations. For each evaluation, we also read the statement of work to determine if the contract covered only the evaluation in our sample or if it covered additional activities as well. In some cases, while the contract covered additional activities, the evaluation’s cost was clearly identifiable in a separate line item or invoice. We identified the evaluation start and end dates using the period of performance in the contract or statement of work. In some cases we determined the period of performance using dates in the evaluation report if the contract’s period of performance covered a broader time period than the evaluation in our sample.

For six USAID evaluations and nine State Department evaluations, we determined the evaluation cost using data from the Federal Procurement Data System – Next Generation (FPDS-NG). FPDS-NG provides a contract’s obligations, start and end dates, and other descriptive data. In each case, we confirmed that the contract covered only the evaluation in our sample by reviewing the statement of work or by confirming with agency officials. We used total obligations to determine the evaluation’s cost and also used the date signed and end date listed in FPDS-NG to determine the period of performance. To assess the reliability of the FPDS-NG data, we (1) reviewed related documentation, (2) traced to or from source documents, and (3) confirmed FPDS-NG data with knowledgeable agency officials. We determined that the FPDS-NG data were sufficiently reliable for the purposes of this engagement.

For MCC, State, and USAID evaluations without clearly identifiable cost information from contract documents or FPDS-NG, we estimated the cost based on budget documents or cost estimates provided by the agency or contractor, where available. We relied on budgets or cost estimates to determine the cost of 5 MCC evaluations, 4 State Department evaluations, and 10 USAID evaluations. We excluded one MCC evaluation from the cost sample because MCC provided a wide range for the estimated cost, and we concluded that this range was not sufficiently reliable to report. We could not determine the cost of one State Department evaluation and one USAID evaluation that were each procured under large agreements that did not separately track evaluation costs. Additionally, USAID officials did not provide cost information for one evaluation.

We report only limited data on the cost of DOD's GT&E and HHS's PEPFAR evaluations because the evaluation contracts or implementing partner agreements did not separately track evaluation costs, and we concluded that the available estimates were too limited to include in our statistical analysis. To estimate the cost of DOD's GT&E evaluations, we reviewed the associated contract and invoices, which included the evaluations as well as additional services. Since the contract and related documents did not contain a separate line item for the evaluations, we requested a cost estimate from agency officials and the contractor. The contractor was able to provide only the broad estimate that we include in our report with appropriate caveats but which we concluded was not sufficiently reliable to include in our statistical analysis. The costs of the HHS PEPFAR evaluations were also not separately tracked by the agency and implementing partners. Evaluation costs were instead estimated by HHS country teams or implementing partners based on their review of previous years' financial records, budgets, or cooperative agreements. Because of the volume of records involved, we judgmentally selected a subsample of 10 HHS evaluations to review the cost estimates provided by HHS officials. To review these estimates, we traced the estimates to source documentation and spoke with knowledgeable agency officials to understand the methodology used to prepare the source estimates. Because of the uncertainty of these cost estimates, we include them in our report with appropriate caveats but concluded that they were not sufficiently reliable to include in our statistical analysis.

To determine the factors that are associated with the costs of foreign aid evaluations, we analyzed the costs of MCC, State, USDA, and USAID evaluations in relation to the data that we collected on these evaluations' quality scores, duration, and other characteristics. We then produced

summary statistics showing the cost differences of various characteristics. For example, we compared the average cost of evaluations with a survey to those without surveys. We conducted difference-in-means tests to determine if any of the characteristics were statistically significant at the 95-percent confidence level and reported characteristics that were significantly related to costs. We also reviewed the evaluations to obtain insights into other likely cost factors, such as unstable locations and the number of sites, for which systematic data were not available for difference-in-means tests. We included location among the characteristics we considered after observing that evaluations that were more costly than others that were of the same type, or required the same performance period to complete, tended to be conducted in unstable or multiple locations.

To assess our third objective, we identified leading practices for the dissemination of evaluation findings. We identified these leading practices using federal guidance, including the President's *Open Government Directive*⁷ and Office of Management and Budget (OMB) guidance,⁸ which encourages or requires the timely public posting of agency information on a searchable website, as well as plans and additional efforts to actively disseminate agency information. In addition to the federal guidance, we also used the American Evaluation Association's (AEA) *An Evaluation Roadmap for a More Effective Government* (AEA Roadmap);⁹ the Organization for Economic Co-operation and Development, Development Assistance Committee's (OECD DAC) *Quality Standards for Development Evaluation*¹⁰ and *Evaluating Development Activities: 12 Lessons from the OECD DAC*;¹¹ and HHS and the General Services Administration's (GSA) *Research-Based Web*

⁷Executive Office of the President, *Open Government Directive*, Memorandum for the Heads of Executive Departments and Agencies (Washington, D.C.: Dec. 8, 2009).

⁸OMB Circular No. A-11, *Preparation, Submission, and Execution of the Budget*, (Washington, D.C.: June 2015).

⁹American Evaluation Association, *An Evaluation Roadmap for a More Effective Government* (October 2013).

¹⁰Organization for Economic Co-operation and Development (OECD), *Quality Standards for Development Evaluation* (Paris, France: 2010).

¹¹Organization for Economic Co-operation and Development (OECD), *Evaluating Development Activities: 12 Lessons from the OECD DAC* (Paris, France: 2013).

*Design and Usability Guidelines*¹² that cite timely public posting, dissemination planning, and additional active efforts to disseminate results as important communication tools for evaluations. We used these sources to identify six practices that agencies should use in order to successfully disseminate the results of foreign aid evaluations. We reviewed the dissemination of all evaluations from fiscal year 2015 for five of the agencies and a sample of USAID evaluations. In total, we examined the dissemination of 193 evaluations: 4 at DOD, 49 at HHS, 16 at MCC, 23 at State, 63 at USAID, and 38 at USDA.

To assess the availability and timeliness of the dissemination of evaluation reports, we reviewed agency policies and websites and interviewed agency officials. We reviewed agency evaluation websites to determine if the evaluation reports in agency evaluation lists had been publicly posted. If an evaluation report had not been posted, we followed up with agency officials regarding the reasons it had not been. We also reviewed the evaluation reports to ensure the documents contained the information necessary for a user to determine if the findings were valid. For example, we reviewed evaluations to ensure that any related annexes had been included when the document had been posted. We examined each agency website to determine whether it provided a search engine that could be used to locate evaluations. We also checked whether the search engine included additional search filters such as the year the evaluation was completed or its location. To assess timeliness, we reviewed agency policies and guidance to determine how soon it required evaluation reports to be posted after the completion of the report. We compared the date an evaluation was considered complete by the agency to the date that it was posted online to determine whether it had been posted within the timeframe required by the agency. To determine whether sensitive evaluations were made available to identified stakeholders via an internal digital system, we reviewed agency lists of sensitive evaluations and interviewed agency officials about agency processes for making sensitive evaluations available internally. We also received an in-person demonstration of the internal posting of USAID's sensitive evaluations and documentation of the internal systems that DOD and USDA use to post evaluations.

¹²HHS and GSA, *Research-Based Web Design and Usability Guidelines*, accessed November 30, 2016, https://www.usability.gov/sites/default/files/documents/guidelines_book.pdf.

To assess agency dissemination planning and its use of additional means for dissemination, we interviewed agency officials and reviewed agency policies, practices, and evaluation documents. To determine if the agency required dissemination planning, we reviewed the dissemination requirements in its evaluation guidance. If the agency required dissemination plans, we reviewed its evaluation reports, contracts, and related documents to determine if they included an identification of the potential users of an evaluation and a description of the approach that will provide users with the evaluation results. If the agency guidance did not require dissemination plans, we asked the agencies if dissemination planning had occurred without the policy in place. If such ad hoc planning had occurred, we asked that the agencies provide examples. We also provided written questions to agency officials regarding additional agency practices for disseminating evaluations other than posting the evaluation online. If agency officials identified additional means of dissemination, we reviewed additional documentary evidence that evaluation findings had been disseminated using these means.

We conducted this performance audit from October 2015 to March 2017 in accordance with generally accepted government auditing standards. Those standards require that we plan and perform the audit to obtain sufficient, appropriate evidence to provide a reasonable basis for our findings and conclusions based on our audit objectives. We believe that the evidence obtained provides a reasonable basis for our findings and conclusions based on our audit objectives.

Appendix II: Evaluation Review Data, by Agency

Agencies varied in the extent to which they met the applicable quality criteria for evaluations that we identified. Tables 9 through 26 below provide further detail on the characteristics and quality of the design, implementation, and conclusions of fiscal year 2015 evaluations we reviewed summarized for all six agencies and then individually for (1) the President's Emergency Plan for AIDS Relief (PEPFAR) programs implemented by the Centers for Disease Control and Prevention (CDC) of the Department of Health and Human Services (HHS), (2) the Millennium Challenge Corporation (MCC), (3) the Department of State (State), (4) the U.S. Agency for International Development (USAID), and (5) the U.S. Department of Agriculture's (USDA) Foreign Agricultural Service's food aid programs.¹

¹Because of the relatively small number of Department of Defense (DOD) evaluations (four evaluations) in our study, we do not include a detailed breakout of the DOD information by question.

Appendix II: Evaluation Review Data, by Agency

Table 9: Quality and Characteristics of the Design of Foreign Assistance Evaluations

Question	Response percentage				Number	Maximum confidence interval
Do the study questions align with the key stated goal(s) of the intervention?	Generally: 94 percent	Partially: 4 percent	Insufficient information: 2 percent	Not at all: 0 percent	173	+/- 3.2 percent
Are the chosen indicators/measures appropriate for the study objectives?	Generally: 83 percent	Partially: 13 percent	Insufficient information: 3 percent	Not at all: 0 percent	173	+/- 5.5 percent
Is the evaluation design appropriate given the study questions?	Generally: 78 percent	Partially: 22 percent	Insufficient information: 0 percent	Not at all: 0 percent	173	+/- 5.9 percent
Does the evaluation assess any net impacts?	Yes: 18 percent	No: 81 percent	Insufficient information: 0 percent		173	+/- 0.4 percent
Does the evaluation state baselines?	Yes: 51 percent	No: 47 percent	Insufficient information: 2 percent		170	+/- 7.2 percent
Does the evaluation state specific targets?	Yes: 54 percent	No: 46 percent	Insufficient information: 1 percent		170	+/- 6.9 percent
Does the evaluation assess processes such as program implementation?	Yes: 88 percent	No: 12 percent	Insufficient information: 0 percent		173	+/- 4.2 percent
Does the evaluation establish criteria such as established plans, budgets, timeframes, and targets?	Yes: 52 percent	No: 48 percent	Insufficient information: 0 percent		149	+/- 7.8 percent
Is the study performed by a third-party evaluator?	Yes: 80 percent	No: 13 percent	Insufficient information: 6 percent		173	+/- 3.8 percent
Are potential conflicts of interest discussed?	Yes: 38 percent	No: 62 percent	Insufficient information: 0 percent		173	+/- 7.1 percent

Source: GAO analysis of fiscal year 2015 foreign assistance evaluation reports for the Global Train and Equip program, administered by the Department of Defense, the President's Emergency Plan for AIDS Relief (PEPFAR) programs implemented by the Centers for Disease Control and Prevention of the Department of Health and Human Services, the Millennium Challenge Corporation, the Department of State, the Department of Agriculture's Foreign Agricultural Service's food assistance programs, and the U.S. Agency for International Development. | GAO-17-316

Notes: Rows may not add to 100 percent due to rounding. Percentages for all agencies combined are weighted to reflect the evaluations in the population that were not selected for the sample.

Appendix II: Evaluation Review Data, by Agency

Table 10: Quality and Characteristics of the Implementation of Foreign Assistance Evaluations

Question	Response percentage				Number	Maximum confidence interval
Are the target population and sampling for the evaluation appropriate for the study questions?	Generally: 56 percent	Partially: 32 percent	Insufficient information: 11 percent	Not at all: 1 percent	173	+/- 7.3 percent
Does the sampling frame appear appropriate?	Generally: 53 percent	Partially: 22 percent	Insufficient information: 24 percent	Not at all: 1 percent	164	+/- 7.5 percent
Is random sampling used?	Yes: 34 percent	No: 44 percent	Insufficient information: 19 percent	Not applicable: 3 percent	173	+/- 6.1 percent
Are the (random) sampling methods appropriate?	Generally: 82 percent	Partially: 7 percent	Insufficient information: 11 percent	Not at all: 0 percent	69	+/- 10.2 percent
Is nonrandom sampling used?	Yes: 70 percent	No: 7 percent	Insufficient information: 21 percent	Not applicable: 2 percent	173	+/- 6.3 percent
Are the sampling methods for nonrandom sampling appropriate?	Generally: 51 percent	Partially: 25 percent	Insufficient information: 23 percent	Not at all: 2 percent	114	+/- 9.1 percent
Is data collection appropriate for the study questions?	Generally: 62 percent	Partially: 32 percent	Insufficient information: 6 percent	Not at all: 0 percent	173	+/- 7.1 percent
Are the data collection methods specified for each question?	Generally: 75 percent	Partially: 15 percent	Insufficient information: 6 percent	Not at all: 4 percent	173	+/- 6.4 percent
Do data collection procedures appear to ensure the reliability of the data?	Generally: 40 percent	Partially: 24 percent	Insufficient information: 24 percent	Not at all: 13 percent	173	+/- 7.0 percent
Does the data analysis appear appropriate to the task?	Generally: 62 percent	Partially: 22 percent	Insufficient information: 16 percent	Not at all: 0 percent	173	+/- 7.1 percent
Are data analysis methods clearly specified for each question?	Generally: 61 percent	Partially: 15 percent	Insufficient information: 9 percent	Not at all: 15 percent	173	+/- 7.2 percent

Source: GAO analysis of fiscal year 2015 foreign assistance evaluation reports for the Global Train and Equip program, administered by the Department of Defense, the President's Emergency Plan for AIDS Relief (PEPFAR) programs implemented by the Centers for Disease Control and Prevention of the Department of Health and Human Services, the Millennium Challenge Corporation, the Department of State, the U.S. Department of Agriculture's Foreign Agricultural Service's food assistance programs, and the U.S. Agency for International Development. | GAO-17-316

Notes: Rows may not add to 100 percent due to rounding. Percentages are weighted to reflect the evaluations in the population that were not selected for the sample.

Table 11: Quality and Characteristics of the Conclusions of Foreign Assistance Evaluations

Question	Response percentage				Number	Maximum confidence interval
Are conclusions supported by the available evidence?	Generally: 68 percent	Partially: 28 percent	Insufficient information: 3 percent	Not at all: 0 percent	173	+/- 6.6 percent
Are the main study limitations identified (in design, data collection methods, and analysis)?	Generally: 56 percent	Partially: 29 percent	Insufficient information: 0 percent	Not at all: 15 percent	173	+/- 7.1 percent
Does the evaluation consider the possibility of unintended side effects of the intervention(s)?	Yes: 28 percent	No: 72 percent	Insufficient information: 0 percent		173	+/- 6.4 percent
Do the evaluation recommendations appear justified by the available evidence?	Generally: 74 percent	Partially: 26 percent	Insufficient information: 0 percent	Not at all: 0 percent	149	+/- 7.3 percent
Are the lessons learned justified by the available evidence?	Generally: 83 percent	Partially: 17 percent	Insufficient information: 0 percent	Not at all: 0 percent	119	+/- 6.8 percent

Source: GAO analysis of fiscal year 2015 foreign assistance evaluation reports for the Global Train and Equip program, administered by the Department of Defense, the President's Emergency Plan for AIDS Relief (PEPFAR) programs implemented by the Centers for Disease Control and Prevention of the Department of Health and Human Services, the Millennium Challenge Corporation, the Department of State, the Department of Agriculture's Foreign Agricultural Service's food assistance programs, and the U.S. Agency for International Development. | GAO-17-316

Notes: Rows may not add to 100 percent due to rounding. Percentages are weighted to reflect the evaluations in the population that were not selected for the sample.

Table 12: Quality and Characteristics of the Design of HHS PEPFAR Evaluations

Question	Response percentage				Number	Maximum confidence interval
Do the study questions align with the key stated goal(s) of the intervention?	Generally: 85 percent	Partially: 12 percent	Insufficient information: 3 percent	Not at all: 0 percent	34	+/- 8.6 percent
Are the chosen indicators/measures appropriate for the study objectives?	Generally: 97 percent	Partially: 0 percent	Insufficient information: 3 percent	Not at all: 0 percent	34	+/- 4.2 percent
Is the evaluation design appropriate given the study questions?	Generally: 79 percent	Partially: 21 percent	Insufficient information: 0 percent	Not at all: 0 percent	34	+/- 8.6 percent
Does the evaluation assess any net impacts?	Yes: 29 percent	No: 71 percent	Insufficient information: 0 percent		34	Not applicable
Does the evaluation state baselines?	Yes: 62 percent	No: 38 percent	Insufficient information: 0 percent		34	+/- 10.4 percent

Appendix II: Evaluation Review Data, by Agency

Question	Response percentage			Number	Maximum confidence interval
Does the evaluation state specific targets?	Yes: 21 percent	No: 79 percent	Insufficient information: 0 percent	34	+/- 7.9 percent
Does the evaluation assess processes such as program implementation?	Yes: 65 percent	No: 32 percent	Insufficient information: 3 percent	34	+/- 9.2 percent
Does the evaluation establish criteria such as established plans, budgets, timeframes, and targets?	Yes: 14 percent	No: 86 percent	Insufficient information: 0 percent	22	+/- 10.6 percent
Is the study performed by a third-party evaluator?	Yes: 26 percent	No: 62 percent	Insufficient information: 12 percent	34	+/- 10 percent
Are potential conflicts of interest discussed?	Yes: 71 percent	No: 29 percent	Insufficient information: 0 percent	34	+/- 10 percent

Legend: HHS = Department of Health and Human Services, PEPFAR = President's Emergency Plan for AIDS Relief

Source: GAO analysis of fiscal year 2015 PEPFAR evaluation reports. | GAO-17-316

Table 13: Quality and Characteristics of the Implementation of HHS PEPFAR Evaluations

Question	Response percentage				Number	Maximum confidence interval
Are the target population and sampling for the evaluation appropriate for the study questions?	Generally: 62 percent	Partially: 21 percent	Insufficient information: 18 percent	Not at all: 0 percent	34	+/- 10.4 percent
Does the sampling frame appear appropriate?	Generally: 54 percent	Partially: 11 percent	Insufficient information: 36 percent	Not at all: 0 percent	28	+/- 11.7 percent
Is random sampling used?	Yes: 29 percent	No: 35 percent	Insufficient information: 18 percent	Not applicable: 18 percent	34	+/- 10.4 percent
Are the (random) sampling methods appropriate?	Generally: 100 percent	Partially: 0 percent	Insufficient information: 0 percent	Not at all: 0 percent	10	+/- 0 percent
Is nonrandom sampling used?	Yes: 38 percent	No: 24 percent	Insufficient information: 24 percent	Not applicable: 15 percent	34	+/- 10.5 percent
Are the sampling methods for nonrandom sampling appropriate?	Generally: 62 percent	Partially: 23 percent	Insufficient information: 15 percent	Not at all: 0 percent	13	+/- 19.2 percent

Appendix II: Evaluation Review Data, by Agency

Question	Response percentage				Number	Maximum confidence interval
Is data collection appropriate for the study questions?	Generally: 76 percent	Partially: 18 percent	Insufficient information: 6 percent	Not at all: 0 percent	34	+/- 9.6 percent
Are the data collection methods specified for each question?	Generally: 82 percent	Partially: 15 percent	Insufficient information: 0 percent	Not at all: 3 percent	34	+/- 7.9 percent
Do data collection procedures appear to ensure the reliability of the data?	Generally: 47 percent	Partially: 15 percent	Insufficient information: 26 percent	Not at all: 12 percent	34	+/- 10.5 percent
Does the data analysis appear appropriate to the task?	Generally: 74 percent	Partially: 18 percent	Insufficient information: 9 percent	Not at all: 0 percent	34	+/- 9.6 percent
Are data analysis methods clearly specified for each question?	Generally: 79 percent	Partially: 12 percent	Insufficient information: 3 percent	Not at all: 6 percent	34	+/- 8.6 percent

Legend: HHS = Department of Health and Human Services, PEPFAR = President's Emergency Plan for AIDS Relief

Source: GAO analysis of fiscal year 2015 PEPFAR evaluation reports. | GAO-17-316

Note: Rows may not add to 100 percent due to rounding.

Table 14: Quality and Characteristics of the Conclusions of HHS PEPFAR Evaluations

Question	Response percentage				Number	Maximum confidence interval
Are conclusions supported by the available evidence?	Generally: 65 percent	Partially: 35 percent	Insufficient information: 0 percent	Not at all: 0 percent	34	+/- 10 percent
Are the main study limitations identified (in design, data collection methods, and analysis)?	Generally: 47 percent	Partially: 38 percent	Insufficient information: 0 percent	Not at all: 15 percent	34	+/- 10.4 percent
Does the evaluation consider the possibility of unintended side effects of the intervention(s)?	Yes: 18 percent	No: 82 percent	Insufficient information: 0 percent		34	+/- 7.9 percent
Do the evaluation recommendations appear justified by the available evidence?	Generally: 81 percent	Partially: 19 percent	Insufficient information: 0 percent	Not at all: 0 percent	27	+/- 10.3 percent
Are the lessons learned justified by the available evidence?	Generally: 76 percent	Partially: 24 percent	Insufficient information: 0 percent	Not at all: 0 percent	21	+/- 13 percent

Legend: HHS = Department of Health and Human Services, PEPFAR = President's Emergency Plan for AIDS Relief

Source: GAO analysis of fiscal year 2015 PEPFAR evaluation reports. | GAO-17-316

Appendix II: Evaluation Review Data, by Agency

Table 15: Quality and Characteristics of the Design of MCC Evaluations

Question	Response percentage				Number
Do the study questions align with the key stated goal(s) of the intervention?	Generally: 94 percent	Partially: 6 percent	Insufficient information: 0 percent	Not at all: 0 percent	16
Are the chosen indicators/measures appropriate for the study objectives?	Generally: 88 percent	Partially: 13 percent	Insufficient information: 0 percent	Not at all: 0 percent	16
Is the evaluation design appropriate given the study questions?	Generally: 88 percent	Partially: 13 percent	Insufficient information: 0 percent	Not at all: 0 percent	16
Does the evaluation assess any net impacts?	Yes: 63 percent	No: 38 percent	Insufficient information: 0 percent		16
Does the evaluation state baselines?	Yes: 75 percent	No: 25 percent	Insufficient information: 0 percent		16
Does the evaluation state specific targets?	Yes: 50 percent	No: 44 percent	Insufficient information: 6 percent		16
Does the evaluation assess processes such as program implementation?	Yes: 75 percent	No: 25 percent	Insufficient information: 0 percent		16
Does the evaluation establish criteria such as established plans, budgets, timeframes, and targets?	Yes: 75 percent	No: 25 percent	Insufficient information: 0 percent		12
Is the study performed by a third-party evaluator?	Yes: 69 percent	No: 0 percent	Insufficient information: 31 percent		16
Are potential conflicts of interest discussed?	Yes: 0 percent	No: 100 percent	Insufficient information: 0 percent		16

Legend: MCC = Millennium Challenge Corporation

Source: GAO analysis of fiscal year 2015 MCC evaluation reports. | GAO-17-316

Note: Rows may not add to 100 percent due to rounding.

Appendix II: Evaluation Review Data, by Agency

Table 16: Quality and Characteristics of the Implementation of MCC Evaluations

Question	Response percentage				Number
Are the target population and sampling for the evaluation appropriate for the study questions?	Generally: 63 percent	Partially: 31 percent	Insufficient information: 6 percent	Not at all: 0 percent	16
Does the sampling frame appear appropriate?	Generally: 73 percent	Partially: 13 percent	Insufficient information: 13 percent	Not at all: 0 percent	15
Is random sampling used?	Yes: 56 percent	No: 19 percent	Insufficient information: 19 percent	Not applicable: 6 percent	15
Are the (random) sampling methods appropriate?	Generally: 100 percent	Partially: 0 percent	Insufficient information: 0 percent	Not at all: 0 percent	9
Is nonrandom sampling used?	Yes: 69 percent	No: 13 percent	Insufficient information: 13 percent	Not applicable: 6 percent	15
Are the sampling methods for nonrandom sampling appropriate?	Generally: 46 percent	Partially: 27 percent	Insufficient information: 27 percent	Not at all: 0 percent	11
Is data collection appropriate for the study questions?	Generally: 69 percent	Partially: 31 percent	Insufficient information: 0 percent	Not at all: 0 percent	16
Are the data collection methods specified for each question?	Generally: 94 percent	Partially: 6 percent	Insufficient information: 0 percent	Not at all: 0 percent	16
Do data collection procedures appear to ensure the reliability of the data?	Generally: 56 percent	Partially: 31 percent	Insufficient information: 13 percent	Not at all: 0 percent	16
Does the data analysis appear appropriate to the task?	Generally: 69 percent	Partially: 25 percent	Insufficient information: 6 percent	Not at all: 0 percent	16
Are data analysis methods clearly specified for each question?	Generally: 81 percent	Partially: 6 percent	Insufficient information: 6 percent	Not at all: 6 percent	16

Legend: MCC = Millennium Challenge Corporation

Source: GAO analysis of fiscal year 2015 MCC evaluation reports. | GAO-17-316

Note: Rows may not add to 100 percent due to rounding.

Appendix II: Evaluation Review Data, by Agency

Table 17: Quality and Characteristics of the Conclusions of MCC Evaluations

Question	Response percentage				Number
Are conclusions supported by the available evidence?	Generally: 75 percent	Partially: 19 percent	Insufficient information: 6 percent	Not at all: 0 percent	16
Are the main study limitations identified (in design, data collection methods, and analysis)?	Generally: 69 percent	Partially: 25 percent	Insufficient information: 0 percent	Not at all: 6 percent	16
Does the evaluation consider the possibility of unintended side effects of the intervention(s)?	Yes: 44 percent	No: 56 percent	Insufficient information: 0 percent		16
Do the evaluation recommendations appear justified by the available evidence?	Generally: 90 percent	Partially: 10 percent	Insufficient information: 0 percent	Not at all: 0 percent	10
Are the lessons learned justified by the available evidence?	Generally: 83 percent	Partially: 17 percent	Insufficient information: 0 percent	Not at all: 0 percent	12

Legend: MCC = Millennium Challenge Corporation

Source: GAO analysis of fiscal year 2015 MCC evaluation reports. | GAO-17-316

Table 18: Quality and Characteristics of the Design of State Evaluations

Question	Response percentage				Number
Do the study questions align with the key stated goal(s) of the intervention?	Generally: 96 percent	Partially: 4 percent	Insufficient information: 0 percent	Not at all: 0 percent	23
Are the chosen indicators/measures appropriate for the study objectives?	Generally: 48 percent	Partially: 43 percent	Insufficient information: 4 percent	Not at all: 4 percent	23
Is the evaluation design appropriate given the study questions?	Generally: 57 percent	Partially: 43 percent	Insufficient information: 0 percent	Not at all: 0 percent	23
Does the evaluation assess any net impacts?	Yes: 9 percent	No: 91 percent	Insufficient information: 0 percent		23
Does the evaluation state baselines?	Yes: 27 percent	No: 73 percent	Insufficient information: 0 percent		22
Does the evaluation state specific targets?	Yes: 5 percent	No: 95 percent	Insufficient information: 0 percent		22
Does the evaluation assess processes such as program implementation?	Yes: 96 percent	No: 4 percent	Insufficient information: 0 percent		23

Appendix II: Evaluation Review Data, by Agency

Question	Response percentage			Number
Does the evaluation establish criteria such as established plans, budgets, timeframes, and targets?	Yes: 27 percent	No: 68 percent	Insufficient information: 5 percent	22
Is the study performed by a third party evaluator?	Yes: 96 percent	No: 0 percent	Insufficient information: 4 percent	23
Are potential conflicts of interest discussed?	Yes: 26 percent	No: 74 percent	Insufficient information: 0 percent	23

Legend: State = Department of State

Source: GAO analysis of fiscal year 2015 State evaluation reports. | GAO-17-316

Note: Rows may not add to 100 percent due to rounding.

Table 19: Quality and Characteristics of the Implementation of State Evaluations

Question	Response percentage				Number
Are the target population and sampling for the evaluation appropriate for the study questions?	Generally: 43 percent	Partially: 43 percent	Insufficient information: 13 percent	Not at all: 0 percent	23
Does the sampling frame appear appropriate?	Generally: 30 percent	Partially: 26 percent	Insufficient information: 44 percent	Not at all: 0 percent	23
Is random sampling used?	Yes: 17 percent	No: 61 percent	Insufficient information: 22 percent		23
Are the (random) sampling methods appropriate?	Generally: 50 percent	Partially: 50 percent	Insufficient information: 0 percent	Not at all: 0 percent	4
Is nonrandom sampling used?	Yes: 78 percent	No: 9 percent	Insufficient information: 13 percent		23
Are the sampling methods for nonrandom sampling appropriate?	Generally: 28 percent	Partially: 44 percent	Insufficient information: 28 percent	Not at all: 0 percent	18
Is data collection appropriate for the study questions?	Generally: 35 percent	Partially: 61 percent	Insufficient information: 4 percent	Not at all: 0 percent	23
Are the data collection methods specified for each question?	Generally: 70 percent	Partially: 17 percent	Insufficient information: 0 percent	Not at all: 13 percent	23
Do data collection procedures appear to ensure the reliability of the data?	Generally: 17 percent	Partially: 35 percent	Insufficient information: 13 percent	Not at all: 35 percent	23

Appendix II: Evaluation Review Data, by Agency

Question	Response percentage				Number
Does the data analysis appear appropriate to the task?	Generally: 48 percent	Partially: 26 percent	Insufficient information: 22 percent	Not at all: 4 percent	23
Are data analysis methods clearly specified for each question?	Generally: 52 percent	Partially: 13 percent	Insufficient information: 13 percent	Not at all: 22 percent	23

Legend: State = Department of State

Source: GAO analysis of fiscal year 2015 State evaluation reports. | GAO-17-316

Note: Rows may not add to 100 percent due to rounding.

Table 20: Quality and Characteristics of the Conclusions of State Evaluations

Question	Response Percentage				Number
Are conclusions supported by the available evidence?	Generally: 61 percent	Partially: 35 percent	Insufficient information: 4 percent	Not at all: 0 percent	23
Are the main study limitations identified (in design, data collection methods, and analysis)?	Generally: 74 percent	Partially: 13 percent	Insufficient information: 0 percent	Not at all: 13 percent	23
Does the evaluation consider the possibility of unintended side effects of the intervention(s)?	Yes: 35 percent	No: 65 percent	Insufficient information: 0 percent		23
Do the evaluation recommendations appear justified by the available evidence?	Generally: 86 percent	Partially: 14 percent	Insufficient information: 0 percent	Not at all: 0 percent	21
Are the lessons learned justified by the available evidence?	Generally: 93 percent	Partially: 7 percent	Insufficient information: 0 percent	Not at all: 0 percent	15

Legend: State = Department of State

Source: GAO analysis of fiscal year 2015 State evaluation reports. | GAO-17-316

Appendix II: Evaluation Review Data, by Agency

Table 21: Quality and Characteristics of the Design of USAID Evaluations

Question	Response percentage				Number	Maximum confidence interval
Do the study questions align with the key stated goal(s) of the intervention?	Generally: 96 percent	Partially: 2 percent	Insufficient information: 2 percent	Not at all: 0 percent	63	+/- 4.6 percent
Are the chosen indicators/measures appropriate for the study objectives?	Generally: 83 percent	Partially: 13 percent	Insufficient information: 4 percent	Not at all: 0 percent	63	+/- 8.6 percent
Is the evaluation design appropriate given the study questions?	Generally: 80 percent	Partially: 20 percent	Insufficient information: 0 percent	Not at all: 0 percent	63	+/- 9.0 percent
Does the evaluation assess any net impacts?	Yes: 14 percent	No: 86 percent	Insufficient information: 0 percent		63	none
Does the evaluation state baselines?	Yes: 45 percent	No: 53 percent	Insufficient information: 2 percent		61	+/- 11.3 percent
Does the evaluation state specific targets?	Yes: 61 percent	No: 39 percent	Insufficient information: 0 percent		61	+/- 10.9 percent
Does the evaluation assess processes such as program implementation?	Yes: 91 percent	No: 9 percent	Insufficient information: 0 percent		63	+/- 6.4 percent
Does the evaluation establish criteria such as established plans, budgets, timeframes, and targets?	Yes: 55 percent	No: 45 percent	Insufficient information: 0 percent		56	+/- 11.8 percent
Is the study performed by a third-party evaluator?	Yes: 91 percent	No: 6 percent	Insufficient information: 3 percent		63	+/- 5.6 percent
Are potential conflicts of interest discussed?	Yes: 42 percent	No: 58 percent	Insufficient information: 0 percent		63	+/- 11.1 percent

Legend: USAID = U.S. Agency for International Development

Source: GAO analysis of fiscal year 2015 USAID evaluation reports. | GAO-17-316

Note: Percentages for USAID are weighted to reflect the evaluations in the population that were not selected for the sample.

Appendix II: Evaluation Review Data, by Agency

Table 22: Quality and Characteristics of the Implementation of USAID Evaluations

Question	Response percentage				Number	Maximum confidence interval
Are the target population and sampling for the evaluation appropriate for the study questions?	Generally: 59 percent	Partially: 33 percent	Insufficient information: 2 percent	Not at all: 6 percent	63	+/- 11.2 percent
Does the sampling frame appear appropriate?	Generally: 54 percent	Partially: 26 percent	Insufficient information: 2 percent	Not at all: 18 percent	62	+/- 11.4 percent
Is random sampling used?	Yes: 31 percent	No: 52 percent	Insufficient information: 18 percent		63	+/- 11.1 percent
Are the (random) sampling methods appropriate?	Generally: 74 percent	Partially: 8 percent	Insufficient information: 19 percent	Not at all: 0 percent	26	+/- 17.8 percent
Is nonrandom sampling used?	Yes: 77 percent	No: 3 percent	Insufficient information: 20 percent		63	+/- 9.7 percent
Are the sampling methods for nonrandom sampling appropriate?	Generally: 54 percent	Partially: 20 percent	Insufficient information: 23 percent	Not at all: 2 percent	48	+/- 12.8 percent
Is data collection appropriate for the study questions?	Generally: 63 percent	Partially: 31 percent	Insufficient information: 6 percent	Not at all: 0 percent	63	+/- 11.0 percent
Are the data collection methods specified for each question?	Generally: 77 percent	Partially: 13 percent	Insufficient information: 6 percent	Not at all: 4 percent	63	+/- 10.0 percent
Do data collection procedures appear to ensure the reliability of the data?	Generally: 43 percent	Partially: 23 percent	Insufficient information: 25 percent	Not at all: 10 percent	63	+/- 10.9 percent
Does the data analysis appear appropriate to the task?	Generally: 63 percent	Partially: 18 percent	Insufficient information: 20 percent	Not at all: 0 percent	63	+/- 10.9 percent
Are data analysis methods clearly specified for each question?	Generally: 58 percent	Partially: 13 percent	Insufficient information: 10 percent	Not at all: 19 percent	63	+/- 11.2 percent

Legend: USAID = U.S. Agency for International Development

Source: GAO analysis of fiscal year 2015 USAID evaluation reports. | GAO-17-316

Notes: Rows may not add to 100 percent due to rounding. Percentages for USAID are weighted to reflect the evaluations in the population that were not selected for the sample.

Table 23: Quality and Characteristics of the Conclusions of USAID Evaluations

Question	Response percentage				Number	Maximum confidence interval
Are conclusions supported by the available evidence?	Generally: 73 percent	Partially: 23 percent	Insufficient information: 4 percent	Not at all: 0 percent	63	+/- 10.2 percent
Are the main study limitations identified (in design, data collection methods, and analysis)?	Generally: 62 percent	Partially: 28 percent	Insufficient information: 0 percent	Not at all: 10 percent	63	+/- 11.0 percent
Does the evaluation consider the possibility of unintended side effects of the intervention(s)?	Yes: 25 percent	No: 75 percent	Insufficient information: 0 percent		63	+/- 10.0 percent
Do the evaluation recommendations appear justified by the available evidence?	Generally: 69 percent	Partially: 31 percent	Insufficient information: 0 percent	Not at all: 0 percent	55	+/- 11.1 percent
Are the lessons learned justified by the available evidence?	Generally: 82 percent	Partially: 18 percent	Insufficient information: 0 percent	Not at all: 0 percent	46	+/- 10.1 percent

Legend: USAID = U.S. Agency for International Development

Source: GAO analysis of fiscal year 2015 USAID evaluation reports. | GAO-17-316

Note: Percentages for USAID are weighted to reflect the evaluations in the population that were not selected for the sample.

Table 24: Quality and Characteristics of the Design of USDA Evaluations

Question	Response percentage				Number	Maximum confidence interval
Do the study questions align with the key stated goal(s) of the intervention?	Generally: 94 percent	Partially: 3 percent	Insufficient information: 3 percent	Not at all: 0 percent	33	+/- 4.7 percent
Are the chosen indicators/measures appropriate for the study objectives?	Generally: 88 percent	Partially: 12 percent	Insufficient information: 0 percent	Not at all: 0 percent	33	+/- 5.7 percent
Is the evaluation design appropriate given the study questions?	Generally: 73 percent	Partially: 27 percent	Insufficient information: 0 percent	Not at all: 0 percent	33	+/- 7.9 percent
Does the evaluation assess any net impacts?	Yes: 18 percent	No: 79 percent	Insufficient information: 3 percent		33	+/- 3.4 percent
Does the evaluation state baselines?	Yes: 82 percent	No: 18 percent	Insufficient information: 0 percent		33	+/- 7.0 percent

Appendix II: Evaluation Review Data, by Agency

Question	Response percentage			Number	Maximum confidence interval
Does the evaluation state specific targets?	Yes: 88 percent	No: 12 percent	Insufficient information: 0 percent	33	+/- 6.4 percent
Does the evaluation assess processes such as program implementation?	Yes: 100 percent	No: 0 percent	Insufficient information: 0 percent	33	none
Does the evaluation establish criteria such as established plans, budgets, timeframes, and targets?	Yes: 70 percent	No: 30 percent	Insufficient information: 0 percent	33	+/- 8.5 percent
Is the study performed by a third-party evaluator?	Yes: 82 percent	No: 9 percent	Insufficient information: 9 percent	33	+/- 6.4 percent
Are potential conflicts of interest discussed?	Yes: 0 percent	No: 100 percent	Insufficient information: 0 percent	33	none

Legend: USDA = U.S. Department of Agriculture

Source: GAO analysis of fiscal year 2015 USDA evaluation reports. | GAO-17-316

Table 25: Quality and Characteristics of the Implementation of USDA Evaluations

Question	Response percentage				Number	Maximum confidence interval
Are the target population and sampling for the evaluation appropriate for the study questions?	Generally: 48 percent	Partially: 24 percent	Insufficient information: 27 percent	Not at all: 0 percent	33	+/- 9.0 percent
Does the sampling frame appear appropriate?	Generally: 55 percent	Partially: 12 percent	Insufficient information: 33 percent	Not at all: 0 percent	33	+/- 9.0 percent
Is random sampling used?	Yes: 61 percent	No: 15 percent	Insufficient information: 24 percent		33	+/- 9.0 percent
Are the (random) sampling methods appropriate?	Generally: 95 percent	Partially: 5 percent	Insufficient information: 0 percent	Not at all: 0 percent	20	+/- 5.7 percent
Is nonrandom sampling used?	Yes: 64 percent	No: 0 percent	Insufficient information: 36 percent		33	+/- 8.9 percent
Are the sampling methods for nonrandom sampling appropriate?	Generally: 48 percent	Partially: 33 percent	Insufficient information: 19 percent	Not at all: 0 percent	21	+/- 11.0 percent

Appendix II: Evaluation Review Data, by Agency

Question	Response percentage				Number	Maximum confidence interval
Is data collection appropriate for the study questions?	Generally: 61 percent	Partially: 30 percent	Insufficient information: 9 percent	Not at all: 0 percent	33	+/- 8.9 percent
Are the data collection methods specified for each question?	Generally: 64 percent	Partially: 21 percent	Insufficient information: 15 percent	Not at all: 0 percent	33	+/- 8.9 percent
Do data collection procedures appear to ensure the reliability of the data?	Generally: 24 percent	Partially: 30 percent	Insufficient information: 27 percent	Not at all: 18 percent	33	+/- 8.5 percent
Does the data analysis appear appropriate to the task?	Generally: 52 percent	Partially: 39 percent	Insufficient information: 9 percent	Not at all: 0 percent	33	+/- 9.0 percent
Are data analysis methods clearly specified for each question?	Generally: 58 percent	Partially: 24 percent	Insufficient information: 12 percent	Not at all: 6 percent	33	+/- 9.0 percent

Legend: USDA = U.S. Department of Agriculture

Source: GAO analysis of fiscal year 2015 USDA evaluation reports. | GAO-17-316

Note: Rows may not add to 100 percent due to rounding.

Table 26: Quality and Characteristics of the Conclusions of USDA Evaluations

Question	Response percentage				Number	Maximum confidence interval
Are conclusions supported by the available evidence?	Generally: 52 percent	Partially: 45 percent	Insufficient information: 3 percent	Not at all: 0 percent	33	+/- 9.0 percent
Are the main study limitations identified (in design, data collection methods, and analysis)?	Generally: 21 percent	Partially: 39 percent	Insufficient information: 0 percent	Not at all: 39 percent	33	+/- 9.0 percent
Does the evaluation consider the possibility of unintended side effects of the intervention(s)?	Yes: 45 percent	No: 55 percent	Insufficient information: 0 percent		33	+/- 9.0 percent
Do the evaluation recommendations appear justified by the available evidence?	Generally: 82 percent	Partially: 18 percent	Insufficient information: 0 percent	Not at all: 0 percent	33	+/- 7.5 percent
Are the lessons learned justified by the available evidence?	Generally: 92 percent	Partially: 8 percent	Insufficient information: 0 percent	Not at all: 0 percent	24	+/- 6.5 percent

Legend: USDA = U.S. Department of Agriculture

Source: GAO analysis of fiscal year 2015 USDA evaluation reports. | GAO-17-316

Note: Rows may not add to 100 percent due to rounding.

Appendix III: Comments from the Department of Defense

Note: GAO comments supplementing those in the report text appear at the end of this appendix.



SPECIAL OPERATIONS /
LOW-INTENSITY CONFLICT

OFFICE OF THE ASSISTANT SECRETARY OF DEFENSE
2500 DEFENSE PENTAGON
WASHINGTON, D.C. 20301-2500

FEB 03 2017

Ms. Jessica A. W. Farb
Acting Director, International Affairs & Trade
U.S. Government Accountability Office
441 G Street, NW
Washington, DC 20548

Ms. Farb,

This is the Department of Defense (DoD) response to the GAO Draft Report, GAO-17-316, "FOREIGN ASSISTANCE: Agencies Can Improve the Quality and Dissemination of Program Evaluations," dated December 28, 2016 (GAO Code 100386).

The Department partially concurs with the recommendation that the "Secretary of Defense...develop a plan for improving the quality of evaluations for the programs included in the review, focusing on areas where our analysis has shown the largest areas for potential improvement." In January 2017, the Department established policy on assessment, monitoring, and evaluation (AM&E) for security cooperation, which will improve the quality of program evaluation across the Department.

The Department will consider improvements consistent with the criteria provided in Tables 9, 10, and 11, Appendix II. In many cases, however, such methodology is not well-suited to support the evaluation of partner security forces. For example, in the context of appropriate target population and sampling methods, it would be unethical for the Department to establish a randomized control group for security assistance evaluation, thus deliberately withholding training or equipment from a group of partner security forces who are fighting alongside or in lieu of U.S. forces. Similarly, evaluator access to partner security forces is often in the context of U.S. assistance and engagement and tied to a shared security mission. Foreign military and security organizations are unlikely to provide significant access to units with whom the U.S. has no partnership simply for the purpose of a U.S. evaluation.

The Department will diligently review the recommendations and best practices among the assessment, monitoring, and evaluation professional community to determine which characteristics are best suited for the unique security sector assistance mission.

See comment 1.

**Appendix III: Comments from the Department
of Defense**

The Department appreciates the opportunity to comment on this draft report. Please direct any questions or comments you may have to Mr. John Raffier, at (703) 614-7055 and john.p.raffier2.civ@mail.mil.



Christopher Maier
Deputy Assistant Secretary of Defense
Special Operations and Combating Terrorism

GAO Comment

DOD partially concurs with our recommendation and notes that in many cases certain methodologies are not well suited for security assistance evaluation. DOD observed that, for example, it would be unethical for DOD to establish a randomized control group for security assistance evaluation and that some foreign military organizations may be unwilling to provide significant access to military units solely for the purpose of an evaluation. We recognize that certain methodologies are not appropriate in every context, and we do not advocate the use of randomized control groups in the evaluations we reviewed for DOD. Our main concerns about the DOD evaluations focus on implementation of the methods used. In particular, we found limitations in sampling methods, including descriptions of the target population; data collection methods; and data analysis. We adjusted pertinent wording in our report to clarify these points.

Appendix IV: Comments from the Department of Health and Human Services



DEPARTMENT OF HEALTH & HUMAN SERVICES

OFFICE OF THE SECRETARY

Assistant Secretary for Legislation
Washington, DC 20201

FEB 02 2017

Jessica A.W. Farb
Acting Director, International Affairs and Trade
U.S. Government Accountability Office
441 G Street NW
Washington, DC 20548

Dear Ms. Farb:

Attached are comments on the U.S. Government Accountability Office's (GAO) report entitled, "*Foreign Assistance: Agencies Can Improve the Quality and Dissemination of Program Evaluations*" (GAO-17-316).

The Department appreciates the opportunity to review this report prior to publication.

Sincerely,

A handwritten signature in cursive script that reads "Barbara Pisaro Clark".

Barbara Pisaro Clark
Acting Assistant Secretary for Legislation

Attachment

**GENERAL COMMENTS OF THE DEPARTMENT OF HEALTH AND HUMAN SERVICES
(HHS) ON THE GOVERNMENT ACCOUNTABILITY OFFICE'S DRAFT REPORT ENTITLED:
FOREIGN ASSISTANCE: AGENCIES CAN IMPROVE THE QUALITY AND DISSEMINATION
OF PROGRAM EVALUATION (GAO-17-316)**

The U.S. Department of Health and Human Services (HHS) appreciates the opportunity from the Government Accountability Office (GAO) to review and comment on this draft report.

Recommendation

In order to better ensure that the evaluation findings reach their intended audiences and are available to facilitate incorporating lessons learned into future program design or budget decisions, GAO recommends that the Secretary of Health and Human Services direct the Centers for Disease Control and Prevention (CDC) to update its guidance and practices on the posting of evaluations to require the President's Emergency Plan for AIDS Relief (PEPFAR) evaluations to be posted within the time frame required by PEPFAR guidance.

HHS Response

HHS concurs with GAO's recommendation. CDC, as of December 2016, now provides guidance noting that each new evaluation protocol specifically state that a report on the main findings of an evaluation will be produced in alignment with the PEPFAR Evaluation Standards of Practice and posted (in English) on a publically accessible website within 90 days of completion. CDC is providing guidance to authors to post the full-length open-access PEPFAR-funded evaluation manuscripts and reports through PubMed or CDC Stacks within the aforementioned 90 days.

Appendix V: Comments from the Millennium Challenge Corporation

Note: GAO comments supplementing those in the report text appear at the end of this appendix.



Date: February 2, 2017

TO: James B. Michels
Assistant Director
International Affairs and Trade
U.S. Government Accountability Office

FROM: Thomas J. Kelly
Acting Vice President
Department of Policy and Evaluation
Millennium Challenge Corporation

SUBJECT: MCC Management Comments on *Foreign Assistance: Agencies Can Improve the Quality and Dissemination of Program Evaluation* (GAO-17-316)

Thank you for the opportunity to review and comment on the U.S. Government Accountability Office's draft report, "Agencies Can Improve the Quality of Dissemination of Program Evaluations." MCC is committed to rigorously evaluating the results of its investments and disseminating the results across the broader development community. MCC welcomes GAO's findings and recommendations for improvement in these areas.

The report examined sixteen independent evaluations that MCC released to the public during Fiscal Year 2015. While MCC is pleased to see that the GAO determined that 88 percent of MCC's evaluations were of "high" or "acceptable" quality, we are unable to provide any comment regarding whether the GAO scored individual evaluations appropriately. The report includes the general methodology for determining "quality," but does not include information on why individual evaluations were categorized. Thus, MCC cannot agree or disagree with any of the GAO's quality determinations.

MCC notes that the report states that "no ... MCC evaluations included [a conflict of interest] statement." Since MCC's adoption of its 2009 Policy on Monitoring and Evaluation of Compacts and Threshold Programs, MCC has required "independent, third-party evaluations" of all MCC projects. In 2013, MCC standardized the language in independent evaluation contracts to explicitly define the following role for evaluators:

Through the evaluation review process, the independent evaluator's role is to assert independence in order to produce a high quality, unbiased evaluation of the program. The independent evaluator should ensure that various stakeholders understand that while MCC, MCA, and other stakeholders may provide guidance and leadership on evaluation questions and feasibility of methodology, once an approved evaluation design is in place, feedback on analysis and interpretation of results is limited to statements on the continued technical and factual accuracy of the analysis.

See comment 1.

See comment 2.

MCC welcomes any guidance regarding conflict-of-interest language that should be used in future independent evaluation reports.

GAO recommends that “the Chief Executive Officer of MCC adjust MCC evaluation practices to make evaluation reports available within the time frame required by MCC guidance.” MCC implemented a stringent review process for all evaluations starting in 2013. The review process ensures that all stakeholders, including partner country governments, have time to comment on the factual accuracy or methodology used in the evaluation report. Both MCC and the partner country government are invited to provide management responses to the evaluation reports, which are released to the public along with the evaluation. When the 2012 Policy on Monitoring and Evaluation was approved, MCC did not anticipate the length of time that the review process would take, as it had not been implemented at the time.

Based on our experience, MCC has revised its Policy on Monitoring and Evaluation of Compacts and Threshold Programs (forthcoming, 2017) to require the following:

All independent evaluation reports are publicly available and posted to the Evaluation Catalog on the MCC website to ensure transparency and accountability. In addition, evaluation reports are accompanied by a summary of findings, which summarizes the key components of the evaluated program, the program logic and accompanying assumptions, monitoring indicators and results, evaluation questions and findings, and key lessons learned by MCC resulting from program implementation and evaluation findings.

Each evaluation has its own Evaluation Catalog entry, which includes a description of methods, key findings, and lessons learned. MCC expects to make each interim and final evaluation report publicly available as soon as practical after receiving the draft report. When applicable, MCC will also publically post statements regarding any significant unresolved differences of opinion between the evaluation and stakeholders.

The Evaluation Catalog also contains microdata generated in the design, implementation, and evaluation of its compacts and threshold programs. All public data sets are approved by MCC’s Disclosure Review Board (DRB), which was established to protect the rights and privacy of individual respondents to MCC-funded surveys. MCC requires public use data files to be free of personal or geographic identifiers that would permit unassisted identification of individual respondents or their household members, and to exclude variables that introduce reasonable risks of deductive disclosure of the identity of individual subjects.

I want to thank you and your staff for the professional manner in which this audit was conducted and for the opportunity to provide additional information and feedback on the GAO draft report. MCC looks forward to continued engagement with GAO to improve its evaluation practices.

Sincerely,

Thomas
Kelly

Thomas Kelly
Acting Vice President
Department of Policy and Evaluation

Digitally signed by Thomas Kelly
DN: cn=Thomas Kelly, o=Millennium
Challenge Corporation, ou=HR,
email=kellyt@mcc.gov, c=US
Date: 2017.02.02 15:36:46 -0500

GAO Comments

1. MCC notes that it has required independent third-party evaluation of all its projects since 2009 and that, in 2013, it standardized the language in its independent evaluation contracts to explicitly define an independent role for evaluators. While these are positive steps, we believe that including in MCC's published evaluations explicit statements about the evaluators' independence and any potential conflicts of interest would bolster the evaluations' credibility and the usefulness.
2. MCC states that it had not anticipated the length of time required by the review process for all evaluations that it implemented beginning in 2013. MCC notes that its forthcoming revised policy on monitoring and evaluation will state that "MCC expects to make each interim and final evaluation report publicly available as soon as practical after receiving the draft report." This revised guidance does not set a specific time frame for the reviews. While agency review efforts may help ensure quality, a specific target for the length of time for the reviews would provide a metric for assessing whether reports are being published in a timely fashion that maximizes their usefulness.

Appendix VI: Comments from the Department of State



United States Department of State
Comptroller
Washington, DC 20520

JAN 27 2017

Charles M. Johnson, Jr.
Managing Director
International Affairs and Trade
Government Accountability Office
441 G Street, N.W.
Washington, D.C. 20548-0001

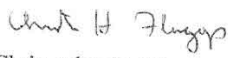
Dear Mr. Johnson:

We appreciate the opportunity to review your draft report, "FOREIGN ASSISTANCE: Agencies Can Improve the Quality and Dissemination of Program Evaluations" GAO Job Code 100386.

The enclosed Department of State comments are provided for incorporation with this letter as an appendix to the final report.

If you have any questions concerning this response, please contact Lisa Crye, Evaluation Advisor, Office of Planning and Systems, Office of Foreign Assistance Resources at (202) 736-4723.

Sincerely,


Christopher H. Flaggs

Enclosure:
As stated

cc: GAO – Jessica A.W. Farb (Acting)
F – Hari Sastry
State/OIG - Norman Brown

Department of State Comments on GAO Draft Report

**FOREIGN ASSISTANCE: Agencies Can Improve the Quality
and Dissemination of Program Evaluations**
(GAO-17-316, GAO Code 100386)

The Department of State welcomes the opportunity to comment on the draft report *Foreign Assistance: Agencies Can Improve the Quality and Dissemination of Program Evaluations*.

Two of the report's recommendations were directed to the Department of State.

The first recommendation directed that each agency covered by the report develop a plan for improving the quality of the evaluations for the programs covered in the review, focusing on areas where GAO analysis showed the largest potential for improvement. We concur with this recommendation and believe that our new *Program Design and Performance Management Policy for Programs, Projects, and Processes*, along with the recently published *Program Design and Performance Management* toolkit and updated policy guidance constitute a plan moving forward.

The second recommendation directed toward the Department of State was to amend the evaluation policy to include a requirement for completion of dissemination plans for each evaluation. We concur with this recommendation and have added language to the draft *Program Design and Performance Management Policy for Programs, Projects, and Processes*, which will be replacing the current evaluation policy.

The Department thanks GAO and the Congress for their efforts to further the use of evaluation and evidence and will use the information contained in this report moving forward.

Appendix VII: Comments from the U.S. Agency for International Development



FEB 02 2017

Ms. Jessica A. W. Farb
Acting Director, International Affairs and Trade
U.S. Government Accountability Office
441 G Street, NW
Washington, DC 20548

Re: FOREIGN ASSISTANCE: Agencies Can Improve the Quality and Dissemination of Program Evaluations (GAO-17-316)

Dear Ms. Farb:

I am pleased to provide the United States Agency for International Development's (USAID) formal response to the U. S. Government Accountability Office (GAO) draft report entitled "*FOREIGN ASSISTANCE: Agencies Can Improve the Quality and Dissemination of Program Evaluations*" (GAO-17-316).

This letter and the enclosed USAID comments are provided for incorporation as an appendix to the final report. Thank you for the opportunity to respond to the GAO draft report and for the courtesies extended by your staff while conducting this GAO engagement.

Sincerely,

A handwritten signature in blue ink, appearing to read "Angelique M. Crumbly".

Angelique M. Crumbly
Acting Assistant Administrator
Bureau for Management

Enclosure: a/s

- 2 -

USAID COMMENTS ON DRAFT REPORT GAO-17-316

USAID appreciates the work undertaken by the GAO, resulting in this report, and the Agency will use the findings to continue to inform how best to target efforts to help staff, external evaluators of USAID programs, and partners understand and meet evaluation quality requirements.

This report has one recommendation for USAID, on page 28-29 of the draft report, as follows:

In order to improve the reliability and usefulness of program evaluations for agency program and budget decisions, we recommend that the Chief Executive Officer of MCC, the Administrator of USAID, the Secretary of Agriculture, the Secretary of Defense, Secretary of State, and Secretary of Health and Human Services (in cooperation with State's Office of the U.S. Global AIDS Coordinator and Health Diplomacy); each develop a plan for improving the quality of evaluations for the programs included in our review, focusing on areas where our analysis has shown the largest areas for potential improvement.

USAID already has a plan for improving the quality of evaluation, including in those areas where this GAO report finds there is the largest potential for improvement: that target populations and sampling are appropriate to the evaluation questions; that data collection methods ensure data reliability; and, that evaluations specify the key assumptions of the data analysis methods used in the evaluation. Steps already taken include (1) recently updating and clarifying the requirements and quality standards for evaluations and (2) working to ensure that staff has the skills they need to manage evaluations through training and other capacity building actions.

Specifically, in September 2016, USAID published the Automated Directives System (ADS) Chapter 201 Program Cycle Operational Policy, with requirements aimed at ensuring high quality evaluations, such as:

- Evaluations will use methods that generate the highest-quality, most credible evidence that corresponds to the questions being asked, taking resources into consideration;
- All required evaluations must be external, i.e. led by an expert external to USAID with appropriate training and experience;
- Draft evaluation statements of work, used to procure external experts, must undergo a peer review;
- All evaluations must have a written evaluation design by the evaluators that describes the key questions, methods, main features of data collection instruments, and data analysis plans. Once final, the design must be shared with implementing partners of the projects or activities addressed in the evaluation and may be shared with other relevant stakeholders;
- USAID staff must plan for dissemination and use of planned evaluations; and
- Evaluation reports must undergo a peer review and should be reviewed against ADS 201maa, Criteria to Ensure the Quality of the Evaluation Report. Reports must also meet the requirements described in ADS 201mah, USAID Evaluation Report Requirements.

- 3 -

To support USAID staff and external evaluators in meeting these requirements and standards, USAID has created and will continue to update online tools and classroom training, and also has evaluation experts on staff to provide direct technical assistance to USAID missions and bureaus at key points in the evaluation planning, design, management, and dissemination processes. For example:

- USAID offers a publicly available “Evaluation Toolkit” which is a resource to guide staff and evaluators through Agency evaluation requirements and best practices; (<https://usaidlearninglab.org/evaluation>).
- USAID offers classroom training in evaluation that more than 1,600 staff members have completed since 2011. This training provides staff with practical skills necessary to better plan for, manage, and utilize the findings from evaluations to inform decisions and learning; and,
- USAID manages a central contract to place fellows with expertise in monitoring, evaluation, learning and project design in USAID missions or offices for six months to two years. Fellows are embedded in technical or program office teams and work alongside them, while sharing expertise and building monitoring and evaluation capacity among their colleagues.

Appendix VIII: Comments from the U.S. Department of Agriculture



United States
Department of
Agriculture

Farm and Foreign
Agricultural
Services

Foreign
Agricultural
Service

1400 Independence
Ave, SW
Stop 1001
Washington, DC
20250-1001

JAN 31 2017

Jessica A.W. Farb
Acting Director, International Affairs and Trade
United States Government Accountability Office
441 G Street, N.W.
Washington, D.C. 20548

Dear Ms. Farb:

The U.S. Department of Agriculture (USDA) appreciates this opportunity to review and comment on the Government Accountability Office (GAO) draft report entitled "Foreign Assistance: Agencies Can Improve the Quality and Dissemination of Program Evaluations" (GAO-17-316). The Department notes GAO's recommendations that it develop a plan for improving the quality of evaluations, and that its Foreign Agricultural Service (FAS) implement guidance and procedures for making FAS evaluations available online and searchable on a single website that can be accessed by the general public. USDA agrees with these recommendations and will take the following actions to address them.

Improve the quality of evaluations

FAS is aware of the importance of quality evaluations. FAS will update its Monitoring and Evaluation Policy for its Office of Capacity Building and Development's Food Assistance Division. Updates will include a requirement to budget 3 – 5 percent of costs for evaluation activities; a requirement to include a separate budget line item for evaluations; and a requirement for evaluators to sign a conflict of interest statement.

FAS also will update its Monitoring and Evaluation Staff Guidance on Reviewing Evaluation Terms of Reference. Updates will include a section on quality that specifically focuses on the following four areas where the GAO analysis has shown the largest areas for potential improvement: 1) ensuring that the target population and sampling for the evaluation are appropriate for the study questions; 2) ensuring that the data collection is appropriate for the study questions; 3) ensuring that the data analysis appears appropriate to the task; and 4) ensuring that conclusions are supported by the available evidence.

Implement guidance and procedures for making FAS evaluations available online and searchable on a single website that can be accessed by the general public

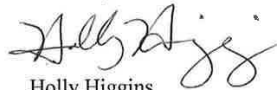
USDA is an Equal Opportunity Employer

**Appendix VIII: Comments from the U.S.
Department of Agriculture**

FAS will continue its current efforts working with its Public Affairs Office to make non-sensitive evaluations publically available online. Once available online, FAS will provide a search engine that can find specific evaluations. Finally, FAS will develop procedures and processes that will establish a timeframe within which non-sensitive evaluations must be made publically available, and develop guidelines for how sensitive evaluations will be made accessible internally.

We would like to thank the GAO for its review and recommendations regarding the quality and dissemination of USDA's foreign assistance program evaluations.

Sincerely,



Holly Higgins
Acting Administrator
Foreign Agricultural Service

Appendix IX: GAO Contact and Staff Acknowledgments

GAO Contact

Jessica Farb (202) 512-6991, or farbj@gao.gov

Staff Acknowledgments

In addition to the contact named above, James Michels, Assistant Director; Thomas Beall, Miranda Berry, Anthony Costulas, Gergana Danailova-Trainor, Martin De Alteriis, Neil Doherty, Mark Dowling, Laurie Ekstrand, Justin Fisher, Georgette Hagans, Kay Halpern, Reid Lowe, Luann Moy, Barry Seltser, Stephanie Shipman, Michael Simon, Douglas Sloane, and Gregory Wilmoth made key contributions to this report.

GAO's Mission

The Government Accountability Office, the audit, evaluation, and investigative arm of Congress, exists to support Congress in meeting its constitutional responsibilities and to help improve the performance and accountability of the federal government for the American people. GAO examines the use of public funds; evaluates federal programs and policies; and provides analyses, recommendations, and other assistance to help Congress make informed oversight, policy, and funding decisions. GAO's commitment to good government is reflected in its core values of accountability, integrity, and reliability.

Obtaining Copies of GAO Reports and Testimony

The fastest and easiest way to obtain copies of GAO documents at no cost is through GAO's website (<http://www.gao.gov>). Each weekday afternoon, GAO posts on its website newly released reports, testimony, and correspondence. To have GAO e-mail you a list of newly posted products, go to <http://www.gao.gov> and select "E-mail Updates."

Order by Phone

The price of each GAO publication reflects GAO's actual cost of production and distribution and depends on the number of pages in the publication and whether the publication is printed in color or black and white. Pricing and ordering information is posted on GAO's website, <http://www.gao.gov/ordering.htm>.

Place orders by calling (202) 512-6000, toll free (866) 801-7077, or TDD (202) 512-2537.

Orders may be paid for using American Express, Discover Card, MasterCard, Visa, check, or money order. Call for additional information.

Connect with GAO

Connect with GAO on [Facebook](#), [Flickr](#), [LinkedIn](#), [Twitter](#), and [YouTube](#). Subscribe to our [RSS Feeds](#) or [E-mail Updates](#). Listen to our [Podcasts](#). Visit GAO on the web at www.gao.gov and read [The Watchblog](#).

To Report Fraud, Waste, and Abuse in Federal Programs

Contact:

Website: <http://www.gao.gov/fraudnet/fraudnet.htm>

E-mail: fraudnet@gao.gov

Automated answering system: (800) 424-5454 or (202) 512-7470

Congressional Relations

Katherine Siggerud, Managing Director, siggerudk@gao.gov, (202) 512-4400, U.S. Government Accountability Office, 441 G Street NW, Room 7125, Washington, DC 20548

Public Affairs

Chuck Young, Managing Director, youngc1@gao.gov, (202) 512-4800, U.S. Government Accountability Office, 441 G Street NW, Room 7149, Washington, DC 20548

Strategic Planning and External Liaison

James-Christian Blockwood, Managing Director, spel@gao.gov, (202) 512-4707, U.S. Government Accountability Office, 441 G Street NW, Room 7814, Washington, DC 20548



Please Print on Recycled Paper.