# FROM EVIDENCE TO LEARNING:

## Recommendations to Improve U.S. Foreign Assistance Evaluation

By
The Lugar Center &
The Modernizing Foreign Assistance Network

the Lugar Center
MFAN MODERNIZING FOREIGN ASSISTANCE NETWORK

# Acknowledgements

The Lugar Center, a 501(c)(3) organization headquartered in Washington, DC under the leadership of Senator Richard G. Lugar (Ret.), is dedicated to proposing solutions to global problems that will define the 21st Century. The Center seeks to educate the public, global policymakers, and future leaders on critical issues such as foreign aid effectiveness, global food security, controlling weapons of mass destruction, and bipartisan governance.

The Modernizing Foreign Assistance Network (MFAN) is a reform coalition composed of international development and foreign policy practitioners, policy advocates, and experts. MFAN was created to build upon the bipartisan consensus that has emerged over the last decade that the U.S. should play a leadership role in achieving economic growth and reducing poverty and suffering around the world and that we can play this role more effectively, efficiently, and transparently.

The Lugar Center and MFAN have joined together to undertake this study of the state of evaluation and learning policies and practices of U.S. foreign assistance in order to inform the new administration, the Congress, and the foreign aid community on ways to ensure continued progress for strengthening the effectiveness of U.S. foreign assistance programs. We engaged independent consultant, Andria Hayes-Birchler, to conduct surveys and interviews that included more than 70 current and former agency staff, foreign aid implementers, evaluators, and civil society partners. Following the initial drafting of the report, we provided separate vetting sessions for stakeholder review. We are grateful for Ms. Hayes-Birchler's work and for the time that the many and diverse stakeholders provided to ensure a highly informed and thorough study.

This report was produced with financial support from the William and Flora Hewlett Foundation.

---

November 2017

# Executive Summary

The United States has been a leader in providing foreign assistance across the developing world for more than 50 years. This foreign policy tool is vital to advancing U.S. interests – promoting security, economic opportunity and our moral values – by helping to ensure that countries can meet the needs of their people and to protect human dignity. While this aid represents only about one percent of the federal budget, it has resulted in the transitioning of some countries from impoverished to middle income, to full trading partners of the United States. In order to ensure that the US Government's (USG) foreign assistance programs are meeting their targets in a cost-effective manner, however, it is vital to conduct and utilize quality evaluations that answer questions such as how funds are spent, whether programs and projects meet their targets, and what the impact is on intended beneficiaries.

Over the past sixteen years, the United States Government has ushered in numerous changes to the evaluation policies and practices of the primary agencies in charge of foreign assistance: The United States Agency for International Development (USAID), the Department of State (State), and the Millennium Challenge Corporation (MCC), as well as the interagency President's Emergency Plan for AIDS Relief (PEPFAR.) Under President Bush, great strides were made to expand and enhance evidence-based foreign assistance through the creation of the MCC and PEPFAR, both of which established clear objectives and benchmarks against which to measure progress. Under President Obama, the primary foreign assistance organizations adopted or revised evaluation policies that outlined specific requirements about when evaluations should be conducted, what types of methodologies are appropriate, who should be responsible, and what level of funding should be allocated to evaluations. Many of these changes aimed to improve the **quantity, quality, and utilization** of evaluations in order to ensure USG foreign aid is as efficient as possible in meeting its objectives.

"Under President Bush, great strides were made to expand and enhance evidence-based foreign assistance through the creation of the MCC and PEPFAR...

Under President Obama, the primary foreign assistance organizations adopted or revised evaluation policies that outlined specific requirements..."

The purpose of this report is to review the state of evaluation policies and practices in agencies implementing USG foreign assistance, as well as provide recommendations to ensure evaluations lead to more evidence-based and effective development policies and programs. This report looks primarily at two types of evaluation: performance evaluations and impact evaluations. When done well, both performance and impact evaluations can help increase **accountability** and **learning** within USG foreign assistance agencies. **Performance evaluations** are often used to answer questions that are pertinent to program design, management, and operational decision-making such as "how was the project implemented?" and "how was the project perceived and valued by various stakeholders?" **Impact evaluations** often seek to test "theories of change" and measure the change in a development outcome that is attributable to a specific intervention (such as reduced incidences of diseases, increased literacy, or increased incomes). Both performance and impact evaluations are vital to understanding whether USG foreign assistance is being delivered effectively and efficiently.

# Evaluations Policies, Practices, Quality, and Utilization

There has been a great deal of progress in evaluation policies and practices related to foreign assistance. In 2011 and 2012, USAID, the State Department, and MCC all adopted or revised evaluation policies that lay out specific requirements and guidance about when evaluations should be conducted, what types of evaluation methodologies are appropriate for various projects, who should be responsible for various evaluation duties, and what type of funding should be allocated to evaluations. Each agency has faced successes and challenges in implementing their policies:

At USAID, the evaluation policy helped increase the number of evaluations from an average of 134 per year from 2006–2010 to an average of 236 per year from 2011–2015.[1] **In the past five years, 1,600 USAID staff have been trained on monitoring and evaluation (M&E),** there are more staff with clearly defined M&E responsibilities, and there is a widely-held perception that the methodologies used in USAID evaluations have become more rigorous and objective. However, many USAID bureaus and missions have struggled to meet the guidance related to funding evaluations, ensuring evaluations of every large project, and overseeing rigorous impact evaluations. After reviewing IG and GAO audits, as well as studies on evaluation quality and evaluation use, USAID updated its evaluation policy in September 2016. The updated policy modified requirements for when evaluations take place, strengthened requirements around evaluation use and dissemination, and encouraged evaluations of higher level outcomes.

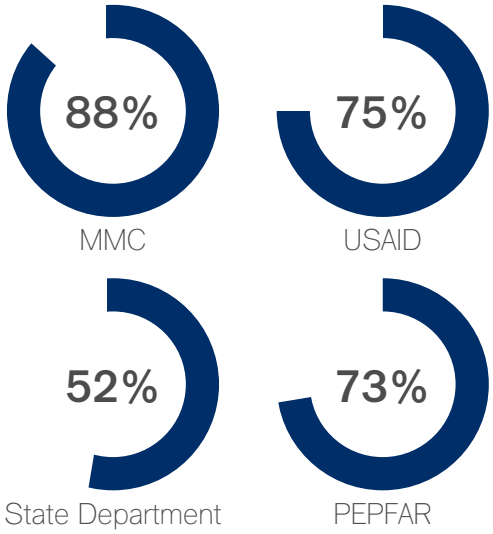
At the State Department, the evaluation policy has raised awareness about the importance of evaluations and built capacity for completing more evaluations. In addition, it has increased the transparency of the State Department's evaluations by requiring non-sensitive evaluations to be posted online within 90 days of completion. However, implementation has been a struggle due to a lack of culture around evaluations, insufficient staff time and capacity, and challenges around evaluating diplomatic, military, or security programs. An audit of State's initial evaluation policy found that 40% of bureaus did not complete the required number of evaluations, and 45% of the documents labeled as 'evaluations' were found not to meet the definition of an evaluation laid out in State's own policy.[2] Since then, **the State Department has used lessons learned to revise its evaluation policy, assist bureaus that struggled to meet the policies' requirements, and develop tools, workshops and consultancies to increase staff's evaluation capacity.**

PEPFAR is an initiative implemented across many agencies under the leadership of the Department of State's Office of the Global AIDS Coordinator and Health Diplomacy. As an interagency initiative, it does not have its own evaluation policy but is subject to the policies of its implementing agencies and organizations. In addition, **it has its own "Evaluation Standards and Practices" report, which lays out evaluation guidance for all its implementing partners.** During the past few years, PEPFAR has made significant strides to ensure more consistent evaluation guidance, standards, and transparency across all its implementing partners. PEPFAR has also invested in timely and comprehensive data collection, including through routine performance data and Population-Based Impact Assessments, improving the possibility that evidence-based decision making can happen in real time.

At MCC, there were strong evaluation practices in place even prior to the revision of its evaluation policy in 2012. According to the Congressional Research Service, "Since its inception, MCC policy has required that every project in a compact be evaluated by independent evaluators, using pre-intervention baseline data. MCC has also put a stronger emphasis on impact evaluation than State and USAID." In 2012, MCC faced its first big test when the first five impact evaluations on farmers' trainings were released. The evaluations showed that MCC met or exceeded its targets on output and outcome indicators, but the results on raising farmers' earnings and incomes were mixed. MCC made a deliberate decision to be as transparent as possible about the impact evaluations and the opportunity for learning they presented. In doing so, **MCC signaled the potential for using evaluations not only to tweak project management but to learn more broadly about what works (and what doesn't) in various sectors.** Since 2012, MCC continues to publish evaluations, summaries of findings, and underlying datasets online. However, it has not been as proactive in making sure evaluations findings are aggregated by sectors, shared broadly with internal or external stakeholders, or used as a 'public good' to proactively inform policy-makers or practitioners.

In 2016, Congress passed the Foreign Aid Transparency and Accountability Act (FATAA). Although FATAA did not create many new requirements or practices for evaluation at the primary aid agencies, it codified many of the reforms the agencies had adopted. It also strengthened the evaluation requirements for many agencies that have smaller foreign assistance portfolios. If implemented and enforced, FATAA has the potential to safeguard the recent improvements in evaluating foreign assistance and serve as a catalyst for further progress. USAID, MCC, and State all updated their evaluation policies explicitly to address FATAA requirements.

Despite progress in improving evaluation policies and practices at USAID, State, MCC, and PEPFAR there are still concerns about the quality of many of their evaluations. The GAO recently found that "medium" or "high quality" evaluations make up 88% of the evaluations at MCC, 75% at USAID, 52% at the State Department, and 73% of the PEPFAR evaluations implemented by the Department of Health and Human Services (HHS.)[3] Generally, these evaluations were asking relevant questions, tracking appropriate indicators, and writing recommendations based on evidence. However, almost 40% of evaluations had inadequate methods for data sampling, collection, and analysis.




"Almost 40% of evaluations had inadequate methods for data sampling, collection, and analysis."

Despite ongoing concerns in regards to the rigor and objectivity of evaluation methods, both quantitative and qualitative data suggest the quality of evaluations has improved in some foreign assistance agencies. Specifically, a meta-evaluation of 37 quality criteria for USAID evaluations showed that 68% of the criteria improved from 2009 to 2012 (before and after the adoption of the evaluation policy.) It also noted that the average composite "quality score" improved from 5.56 to 6.69 (on a scale of zero to ten) during the same time period.[4] Of the criteria that showed notable improvements, three stand out as particularly important:

The percent of USAID evaluations where management's purpose for undertaking the evaluation is explicitly described increased from 70% to 81%. This suggests evaluations are not being done simply out of compliance but due to real interest from management on increasing accountability or learning.

The percent of USAID evaluations where study limitations were included increased from 38% to 64%. This suggests evaluators, agencies, and other stakeholders are paying more attention to methods and being more transparent about methodological limitations.

The percent of USAID evaluations where the recommendations are specific about what is to be done increased from 58% to 77%. This suggests both evaluators and donors are increasingly understanding the importance of making evaluation recommendations useful for future program design or decision-making.

In addition, a number of interview respondents for this assessment commented on the increased quality of scopes of work for evaluations coming out of USAID, higher demand and interest in evaluation methodologies at USAID, more attention to the usefulness of evaluation design at MCC, and increased attention to timeliness of evaluations at MCC.

Progress has been made in making evaluations more accessible and better utilized by development stakeholders. However, **there are still numerous barriers to optimal utilization of evaluation findings and recommendations.** Many interview respondents spoke about the necessary conditions for utilizing evaluations to improve aid effectiveness. They tend to fall into three categories:

Evaluations must be built around rigorous and objective methodologies and deliver credible findings, lessons learned, and recommendations.

Evaluations must be accessible - in both a literal and figurative sense - to a wide array of stakeholders, including program designers, program implementers, policy-makers, and citizens in both the donor and beneficiary countries.

There must be opportunities, incentives, and systems in place to design and modify programs based on evaluation findings.

As explored above, progress has been made on improving the rigor and objectivity of evaluations, although there is still a way to go in improving methods used for data sampling, collection, and analysis, as well as considering external factors that might affect the project's results. In addition, progress has been made to ensure evaluations are accessible. Most non-sensitive evaluations are now posted online in a timely manner, and improvements to evaluation catalogues at USAID and MCC have made evaluations easier to access. In addition, USAID and MCC have both taken steps to proactively share evaluation findings with development policymakers and practitioners by publishing concise summaries of findings, circulating newsletters, hosting evidence summits, and developing online platforms to share evaluation findings and recommendations.

However, there are still a number of barriers in terms of accessibility, opportunities, incentives, and systems in place to properly use evaluations. These include inflexible budgets (and grants or contracts) that do not allow for adaptation based on evaluation findings, lack of staff time to respond effectively and follow-up on evaluation findings or recommendations, lack of demand for evidence-based program design from senior leadership, and lack of incentive to use scarce resources for developing and utilizing evaluations.

# Recommendations

In order to ensure further progress on these issues, the new administration, Congress, and agency leadership should adopt the following recommendations.

## Evaluation Policies

- The new administration should appoint leadership that has demonstrated success using evaluations to inform evidence-based policy-making.
- Congress should provide rigorous oversight of the implementation of FATAA, both as the Office of Management and Budget (OMB) writes the agency guidance and as the agencies carry out this guidance.
- OMB should set a high bar for best practices in evaluation policies when drafting guidance on implementing FATAA. It should ensure its guidelines include not only guidance on when to conduct evaluations, but also guidance on how to ensure evaluations are high quality, and guidance on the dissemination of evaluations that goes beyond posting completed evaluations online. OMB should also ensure that agencies follow their own evaluation policy guidance and set aside resources for evidence-building activities (such as evaluations) in their budget requests.

## Quality of Evaluations

- **Agencies should increase the quality of evaluations by demanding higher standards for data sampling, collection, and analysis,** and by ensuring resources are available to meet these standards. One of the main costs of evaluations is collecting primary data, often through surveys of program recipients and non-recipients. A lack of adequate resources is likely one of the reasons more than 40% of evaluations still do a poor job sampling and collecting data. When possible and relevant evaluators should be encouraged to share and make public their primary data.
- Agencies should consider prioritizing additional impact and ex-post evaluations, when appropriate. While impact and ex-post evaluations are not inherently more rigorous than performance evaluations, they do add distinct value to an overall evaluation portfolio. By prioritizing additional impact and ex-post evaluations when appropriate, agencies can ensure they have a more diverse and robust pool of information to inform their policies and programs.

## Accessibility and Disseminations of Evaluations

- Agencies and evaluators should work to make evaluations more accessible. Specifically, they should increase the publication of summaries of findings and similar concise communications focused on learning and policy actions; ensure wider and better distribution of evaluations and findings; and work to cluster and synthesize evaluations findings by sectors.
- **Agencies and evaluators should include partner governments, local populations, and NGOs in designing, implementing, and responding to evaluations.** Local evaluators are included in only a third of USAID's evaluations, and evaluations are shared with country partners less than a third of the time at USAID. Agencies should consider adopting MCC's model of including local stakeholders and partner governments as a part of the official review process prior to evaluation completion.

## Utilization of Evaluations for Accountability and Learning

- Congress should create more flexible budgets for agencies so that they may utilize evaluation findings to inform resource allocation. If agencies demonstrate that they are, in fact, shifting funds at the project level based on what they have learned through evaluations, Congress should respond, allowing for more flexible budgets so that learning can take place at a larger, program level scale too. This increased flexibility would require a reduction in earmarks, a policy that MFAN has long endorsed. Similarly, when approving projects, agency leadership should request how evaluation findings and evidence have informed project designs. This should be done systematically, not on occasion. Prior to approving any project, leadership should ask its designers, "Have similar projects been done before?" If there are no prior evaluations, leadership should ensure the project has a clear development hypothesis and encourage the design of an evaluation strategy to ensure findings and lessons learned exist next time for similar projects.
- Program and operations staff should be more involved in designing scopes of work for evaluations, as well as the evaluation itself. In addition, there should be staff in agencies, implementing partners, and independent evaluators whose primary – if not sole – responsibility is ensuring evaluation findings are shared, internalized, and acted upon. By ensuring there are full-time staff dedicated to learning and sharing information from evaluations, leadership can both solve a primary constraint to utilizing evaluation findings – the lack of staff time – and also signal the importance their organization puts on evidence-based programming and policy-making.
- Agency leadership should put systems in place to ensure evaluation recommendations are systematically shared with relevant stakeholders and a plan is put in place to respond to recommendations.
- Agency leadership should develop a set of concrete examples for use internally and with Congress and outside stakeholders, in order to demonstrate the critical role that evaluation and learning play in supporting effective foreign assistance.
- OMB should serve as a consumer of evaluation findings by asking agencies how evidence and evaluations inform their budget requests. **By linking both evaluation resources and evaluation findings into the budget build process, OMB can help ensure evaluations are utilized to achieve the most effective outcomes.**

When Congress passed the Foreign Aid Transparency and Accountability Act in 2016, it did so with broad bipartisan support. It was welcomed with equal enthusiasm by the advocacy community. Even staff at the foreign assistance agencies – who often view congressional requirements as burdensome – welcomed FATAA as an important piece of legislation that safeguards the progress made in evaluating foreign assistance.

This widespread, bipartisan support across policymakers, practitioners, and advocates reflects a shared agreement that effective foreign assistance programs strengthen America's standing in the world, bolster its national security, and save or improve the lives of millions of people around the world. However, **a critical component for ensuring this assistance is as effective as possible is by producing and utilizing high-quality evaluations.** This report summarizes the progress made in evaluating foreign assistance, but it also shows that there is still more work that needs to be done to transition agencies from fully implementing their own evaluation policies and adhering to FATAA provisions to developing an ingrained culture that values evaluations as a means for learning and implementing what works.

# 01 Introduction

The United States has been a leader in providing foreign assistance across the developing world for more than 50 years. This foreign policy tool is vital to advancing U.S. interests – promoting security, economic opportunity and our moral values – by helping to ensure that countries can meet the needs of their people and to protect human dignity. While this aid represents only about one percent of the federal budget, it has resulted in the transitioning of some countries from impoverished to middle income, to full trading partners of the United States. In order to ensure that the US Government's (USG) foreign assistance programs are meeting their targets in a cost-effective manner, however, it is vital to conduct and utilize quality evaluations that answer questions such as how funds are spent, whether programs and projects meet their targets, and what the impact is on intended beneficiaries.

Over the past sixteen years, the United States Government has ushered in numerous changes to the evaluation policies and practices of the primary agencies in charge of foreign assistance: The United States Agency for International Development (USAID), the Department of State (State), and the Millennium Challenge Corporation (MCC), as well as the interagency President's Emergency Plan for AIDS Relief (PEPFAR.) **Under President Bush, great strides were made to expand evidence-based foreign assistance through the creation of the MCC and PEPFAR, both of which established clear objectives and benchmarks against which to measure progress. Under President Obama, each of the primary foreign assistance organizations adopted or revised evaluation policies which outlined requirements about when evaluations should be conducted,** what types of evaluation methodologies are appropriate, who should be responsible, and what type of funding should be allocated to evaluations. Many of these changes aimed to improve the quantity, quality, and utilization of evaluations in order to ensure USG foreign aid is as efficient as possible in meetings its objectives.

The current administration has questioned the value of foreign assistance, and one of the best ways to assess and increase the value (including the cost-effectiveness) of US foreign aid is to ensure the completion and utilization of quality evaluations of programs and projects. Recognizing this, Congress passed the Foreign Aid Transparency and Accountability Act (FATAA) in 2016 with broad, bipartisan support. The widespread support of FATAA reflects a recognition that effective foreign assistance programs strengthen America's standing in the world, bolster its national security, and save or improve the lives of millions of people around the world.

This report assesses the progress of evaluation practices in the primary foreign assistance agencies, as well as ongoing barriers, in order to make recommendations on how to best ensure evaluations are rigorous, high quality, and useful in making US foreign assistance as effective as possible. In a time when budget cuts to US foreign assistance seem likely, it is more important than ever to ensure limited funding is utilized as efficiently as possible by the USG and its partners to achieve the greatest impacts.

# Scope, Methodology, and Limitations of the Report

In order to inform recommendations for the new administration and Congress, this report examines the following:

**USG Foreign Assistance Evaluation Policies and Practices:**

This section examines the progress of evaluation policies and practices at USAID, State Department, MCC, and PEPFAR . It highlights the role new or revised evaluation policies have played in strengthening evaluations practices in each agency, as well as the obstacles they have faced during implementation. It also examines the role FATAA has played in improving the evaluation policies and practices in these agencies.

**Quality of evaluations:**

This section lays out various definitions and metrics for assessing the quality of evaluations and presents findings on evaluations' quality at each agency. It looks at progress in evaluation quality over time, comparative quality across agencies, and ongoing issues related to the quality of foreign assistance evaluations.

**Utilization of evaluations.**

This section examines whether and how internal and external stakeholders use evaluations to improve development effectiveness. It looks at the use of evaluations for both accountability and learning. This section also examines barriers to use and what can be done to overcome them.

This assessment builds on the rapidly growing literature on these topics, including two meta-evaluations on the quality and utilization of evaluations at USAID, OIG audits of evaluation policies at USAID and State, lessons learned by MCC on their evaluation practices and findings, blogs and reports from think tanks such as the Center for Global Development (CGD) and civil society advocates such as the Modernizing Foreign Assistance Network (MFAN), and recent publications from the Congressional Research Service (CRS) and Government Accountability Office (GAO) on similar topics. The scope and focus of this report differs somewhat from the aforementioned literature in the following ways: While many of the previous reports or audits focus on only one agency, this report assesses USAID, MCC, the State Department, and PEPFAR. In addition, this report focuses more specifically than the others on barriers to evaluation utilization and how to overcome these barriers.

In addition to the literature review, this assessment uses information from surveys of 70 stakeholders and interviews of over 35 evaluation experts, including staff from the USG and implementing partners, independent evaluators, and civil society advocates. Respondents include current and former staff from the State Department, USAID, MCC, and the House Foreign Affairs Committee, as well as current and former staff from the Modernizing Foreign Assistance Network, the Center for Global Development, Brookings, Oxfam, CARE, Save the Children, Publish What You Fund, Social Impact, Mathematica Policy Research, Management Systems International (MSI), and the National Democratic Institute (NDI.) For the sake of candidness, both survey and interview data remains anonymous. The complete list of interview questions and survey questionnaire can be found in Annexes 1 and 2, respectively.


**Recommendations:**

This section lays out specific recommendations about how to improve the quantity, quality, and utilization of evaluations in USG foreign assistance agencies. It highlights what actions have been most effective in increasing the quality of evaluations and what actions are needed to increase quality further. Similarly, it lays out what actions have been most effective in increasing the use of evaluations and what actions are needed to increase use further.

This report focuses more specifically than the others on barriers to evaluation utilization and how to overcome these barriers.

Given the topic of this report, it is important to note that this report is not itself an evaluation. It is an assessment based on existing literature, as well as the perceptions and experiences of development stakeholders both internal and external to the USG. While the authors did try to consult a wide-breadth of stakeholders, the survey and interview respondents may not be representative of the entire development or evaluation communities. This assessment focuses exclusively on evaluations of USG foreign assistance. It does not examine monitoring systems and tools, audits, or research related to USG foreign assistance. Similarly, it does not assess evaluation policies or practices across non-USG donors, except as they tie back to USG funding or inform recommendations for the USG.

However, it does examine several different purposes and types of evaluations, which are important to define up front. Various stakeholders conceptualize evaluations very differently and, as such, it is necessary to lay out some basic terminology that will be used in this assessment.

# There are two primary purposes for evaluations: accountability and learning.

When many stakeholders discuss the importance of evaluations, they focus primarily on the role of evaluations in ensuring accountability. They want to know where and how the money was spent, whether the project met its objectives, what worked, and what didn't in the project. In general, evaluations focused on accountability measure outputs that are under the direct control of the agency or implementers, so that success or failure of project implementation can be linked to specific groups or individuals (and they can be held accountable).

Other stakeholders are more interested in evaluations for the purpose of learning. They point out that there are many "theories of change" in development that have not actually been rigorously tested. Often, they are interested in evaluations that demonstrate whether a well-designed and well-implemented project is successful in changing specific outcomes or impacts for the intended beneficiaries.

# There are two main categories of evaluations: performance evaluations and impact evaluations.

**Performance evaluations** are often used to answer descriptive and normative questions such as "What did this particular project achieve?" "How was it implemented?" "How was it perceived and valued?" "Are the expected results occurring?" "What are unintended outcomes?" They may include other questions that are pertinent to program design, management, and operational decision making.

For example, a performance evaluation of a literacy project might look at how much money an NGO spent purchasing books for elementary schools, whether the books were successfully delivered to the schools, whether students were using the books once they were delivered, and whether there were unintended outcomes. If the books were never delivered, or some were missing, or they were not in use when the evaluators observed the school, this would suggest there were problems in project implementation that need to be resolved – ideally for the current project and certainly before the next literacy project.

**Impact evaluations** are often used to establish what the impact of the project was on the intended beneficiaries and what the impact might have been without the intervention, in order to control for factors other than the intervention that might account for observed changes (such as changes in government policies, interventions from other donors, weather patterns, disease outbreaks, etc.) This type of evaluation allows the evaluators and other stakeholders to attribute changes in a development outcome to a specific intervention.

Using the above example, it may be the case that even if the books were successfully delivered and used by students, they may not contribute to increased literacy for reasons that have nothing to do with project implementation. It may be the case that distributing books is not actually an effective method of increasing literacy or is not sufficient without other interventions (improved teacher training, translating books into the local language, ensuring proper nutrition of students, etc.) In general, impact evaluations want to test fundamental assumptions about project design or want to prove that a project caused a specific change for its intended beneficiaries.

"It is not the case that an impact evaluation is inherently more rigorous or "better" than a performance evaluation."

"It is the case that only a well-designed and implemented impact evaluation can tell stakeholders whether (or the degree to which) specific outcomes are attributable to an intervention."

It is important to note that both types of evaluations can be useful in fully assessing a development project. In the initial example, an impact evaluation might find that the students' literacy levels did not increase, relative to other students who were not beneficiaries of the project. However, if there was no evaluation of the project's implementation it would be impossible to determine if the lack of results is because the project was poorly implemented or because the theory of change was faulty. Conversely, if a performance evaluation determined the project was perfectly implemented, it still wouldn't tell you whether the students' literacy rate improved as a result of the project.

Although both types of evaluations are useful and necessary, it is vital to be clear about what type of evaluation is being discussed. Different types of evaluations have different costs, methods, purposes, and findings. Many misconceptions about evaluations come through a lack of clarity in language.[5] For example, it is not the case that an impact evaluation is inherently more rigorous or "better" than a performance evaluation. Each type of evaluation has its own set of methodologies (which can be well-designed and implemented or not) and sets out to gather its own type of evidence to inform its findings and recommendations (which can be helpful and informative or not.) However, it is the case that only a well-designed and implemented impact evaluation can tell stakeholders whether (or the degree to which) specific outcomes are attributable to an intervention. For an example of why methods – and clear language around methods - matter, see below.

# *Why Methods Matter:*
## Two Reports on the Same Project Come to Very Different Conclusions

From 2006 to 2008, USAID implemented an MCC threshold program to combat corruption. One of the projects trained journalists on anti-corruption statutes and investigative techniques. According to USAID's final report, "More than 300 journalists were trained in investigative reporting . . . and these journalists have had a significant impact on the media environment in the country. At the inception of the (project), Tanzanian newspapers were publishing only 20 corruption-related stories per month. During the final quarter of the program, this figure reached 428 per month. During the 31 months of the Program, the Tanzanian media published 6,363 stories on corruption, nearly five times the target of 1,300." This sounds like a very successful project: targets were exceeded and progress can be demonstrated over time using before and after data.

However, the data above includes all corruption-related newspaper stories, not just those written by journalists trained by the project. The report does not try to determine the counterfactual (i.e. what would have happened if the journalists had not received training) nor does it examine other possible reasons for a spike in corruption-related stories. However, an independent evaluation of the same project used different methods. It tracked the number of articles written by trained and untrained journalists, both before and after the trainings. It found that on average trained journalists did not increase their rate of publication any more than untrained journalists. In other words, all the journalists were publishing more corruption-related stories in 2008, not just those who received training.

The same independent evaluation found that a separate anti-corruption project in the same threshold program, which focused on improving the government's capacity for auditing procurements, created a "substantial and significant improvement" in procurement compliance. One audit uncovered large irregularities in the procurement of electrical generators by the national electricity purveyor. The cost of the mismanagement exceeded $68 million, and the resultant scandal led to the resignation of the prime minister and several other members of the government.

One might therefore surmise the huge uptick in corruption-related newspaper stories in 2008 was due not to the success of the journalist training but the success of the auditing project. By uncovering the financial irregularities, the auditors unveiled a massive corruption scandal which was covered by all journalists (not just those who received training.)

While there are additional purposes and types of evaluations in foreign assistance, this report focuses primarily on performance and impact evaluations for the purposes of accountability and learning. This is not because these types or purposes are the most important, but simply because these classifications are used by most of the referenced literature and most of the survey and interview respondents. This language also mirrors the terminology used in USAID and MCC's evaluation policies.

That said, it is worth also briefly highlighting the importance of ex-post evaluations for the purpose of evaluating the sustainability of USG foreign assistance. An ex-post evaluation is conducted after a certain period has passed since the completion of a project, usually to assess the sustainability of the project and its impacts over time. These types of evaluations are still relatively rare in USG foreign assistance agencies, although USAID did release a series of ex-post evaluations on its Food for Peace program in 2016. Although infrequently discussed in the evaluation policies, literature, or by TLC/MFAN's survey or interview respondents, ex-post evaluations are critical to ensuring USG foreign assistance is delivering sustainable results.

"A major contribution of the FATAA legislation is that it placed evaluation within the continuum of performance management,"

# Both performance and impact evaluations can lead to better use of development assistance funds.

Without performance evaluations, it is difficult to know whether a project is implemented in a timely and efficient manner, whether output targets are being met, whether beneficiaries value the outputs and outcomes of a project, and whether there were unintended positive or negative consequences of the project. Without impact evaluations, it is difficult to know whether an intervention is actually effective in changing outcomes or impacts (i.e. literacy rates, incidence of water borne diseases, household incomes, etc.) for the intended beneficiaries.

When evaluations are well-designed, implemented, and utilized, they can help development stakeholders have more impact. The influence of evaluations may include modifying project design to avoid previously encountered issues, scaling up a pilot project that appears to be effective, or reducing funding to an intervention that does not appear to be making an impact on its intended beneficiaries.

One example of this is a performance evaluation of a USAID forestry project in Indonesia, which found that the project was spread too thin across too many geographic locations to be very effective. In response, USAID consolidated down from fifteen locations to five and then actually increased its budget in those five locations. "We realized that we were . . . a mile wide and an inch deep and that we really needed to go deeper and have more impactful interventions," said staff in USAID/ Indonesia's Environment Office, "It was only because of that evaluation that we made that decision."[6]

"When evaluations are well-designed, implemented, and utilized, they can help development stakeholders have more impact."

At MCC, an impact evaluation of a water supply project in rural Mozambique tested the assumption that if women and girls spent less time collecting water, they would spend more time earning income or attending school. MCC invested $13 million on 615 improved water sources, benefitting 317,000 people. As a result, women and girls spent one to two fewer hours per trip collecting water (despite collecting 9 – 33% more water.) However, the evaluators found that time savings for women and girls were usually directed to domestic activities, small scale farming, and resting, not income generation or school. This evaluation made MCC rethink how it values "time savings." In the Beijing Platform for Action for the Advancement of Women, development institutions are encouraged to measure "the work done by women and men, including both remunerated and unremunerated work." This evaluation helped MCC recognize that value can be found in non-income generating work, including the domestic activities many of the women and girls prioritized with their newly available time.

These are two of many examples of how evaluations can help the foreign assistance community learn how to have more impact, test assumptions, and revise projects and models accordingly.

# 02 USG Foreign Assistance Evaluation Policies and Practices

In 2009, three evaluators and former USAID employees released a paper entitled "Beyond Success Stories: Monitoring & Evaluation for Foreign Assistance Results."[7] The paper assessed the monitoring and evaluation (M&E) systems and practices at several US government agencies in charge of foreign assistance, primarily USAID, MCC, and the Department of State. The report painted a relatively bleak picture: there were not centralized evaluation policies in place in most of the agencies; evaluation policies, practices, systems, and tools varied dramatically not only across agencies but also within agencies' bureaus, offices, and missions; baseline data to track progress against was rare; funding for evaluations was often limited; the quality of evaluations was often perceived as insufficiently rigorous to provide credible evidence, findings, or recommendations; and even if evaluations were rigorous they were generally not utilized outside of the immediate operating unit or staff that commissioned the evaluation.

This report was not alone in expressing concerns about the lack of rigorous, useful evaluations of USG foreign assistance. The Center for Global Development had previously launched a report entitled "When Will We Ever Learn? Improving Lives through Impact Evaluations, which called on "developing countries, aid agencies, foundations and NGOs to close the evaluation gap by adopting good practices in terms of independently evaluating the impact of their own programs." Meanwhile the Modernizing Foreign Assistance Network (MFAN) was encouraging Congress to pass the Foreign Assistance Revitalization and Accountability Act of 2009, in part to address the lack of planning, policy, and evaluation capability at USAID.

Evaluations of USG foreign assistance appeared to be at its nadir in 2009. The number of evaluations submitted to USAID's Development Experience Clearinghouse (DEC) had decreased from nearly 500 in 1994 to approximately 170 in 2009, despite an almost three-fold increase in program dollars managed by USAID.[8] USAID's Policy Bureau and its Center for Development Information and Evaluation had been eliminated. USG agencies such as the Department of State, the Department of Defense, and others were playing an increasing role in planning and implementing foreign assistance programs. However, very few of these agencies had policies or practices in place to ensure the evaluation of their programs. In addition, funding and staff for evaluations was limited in many foreign assistance agencies.

At MCC, however, there was a different environment. Founded in 2004 under the Bush administration, MCC required independent evaluations of all compact projects, laid out project indicators during compact design, collected baseline data prior to project implementation, and referred to impact evaluations as an 'integral' part of MCC's model. In fact, about 40% of MCC's evaluations were impact evaluations. According to William Savedoff at CGD, the proportion of impact evaluations at MCC was "certainly a record for a development agency, possibly even for most other public and private organizations."[9]

However, the long-term nature of MCC's compacts - which often take two to three years to develop and then five years to implement - meant that as of 2009 most MCC evaluations were still in progress. There were many open questions about whether the impact evaluations would successfully capture the information MCC needed – both in terms of whether the projects were successfully meeting their outcomes and impact targets, and whether there would be sufficient qualitative information to answer why targets were or were not met. While there was enthusiasm for MCC's commitment to evaluation, there was not yet much evidence about whether the investment in evaluations would pay-off.

Within PEPFAR, there was a culture and practice of conducting evaluations, but the quality of the evaluations varied across implementing partners, in part because there was no standard guidance applied to all implementers. When Congress reauthorized PEPFAR in 2008, it mandated the Institute for Medicine (IOM) conduct a systematic evaluation of PEPFAR's overall progress (in addition to the decentralized, program-specific evaluations being conducted at the time.) The legislative requirement seemed to reflect a broader concern about whether development assistance was meeting not just output targets but outcome targets.

President Obama's administration also stressed the importance of tracking outcomes in foreign assistance. In 2009, he encouraged the State Department to embark on a Quadrennial Diplomacy and Development Review (QDDR) to highlight lessons learned and make recommendations for improvements. The QDDR ultimately recommended "focusing on outcomes and impact rather than inputs and outputs, and ensuring that the best available evidence informs program design and execution."[10] The following pages describe how each agency worked to implement this recommendation.

# A. USAID

In June 2010, in a demonstration of senior leadership commitment to evaluation and results, USAID formed a new Office of Learning, Evaluation, and Research in its Bureau for Policy, Planning, and Learning (PPL/LER). Ruth Levine was chosen to lead PPL, oversee the adoption of an Evaluation Policy for USAID, and help rebuild evaluation capacity across the agency and its missions. The choice of Ruth Levine to play this role was significant, as she had previously served on the Center for Global Development's Evaluation Working Group and co-authored "When Will We Ever Learn? Improving Lives through Impact Evaluation." Many staff had concerns about the appropriateness and feasibility of impact evaluations at USAID, ranging from concerns about the costs, the limited capacity of staff to manage such technical evaluations, and the feasibility of designing evaluations that were relevant to the work of USAID (which often times focuses on technical assistance, humanitarian assistance, or working in fragile or conflict areas.)

To address these concerns, PPL ensured the evaluation policy was supportive of many types of evaluations and evaluation methodologies. The policy states: "Given the nature of development activities, both qualitative and quantitative methods yield valuable findings, and a combination of both often is optimal; observational, quasi-experimental and experimental designs all have their place. No single method will be privileged over others; rather, the selection of method or methods for a particular evaluation should principally consider the empirical strength of study design as well as the feasibility."[11] Further guidance on evaluation design and methodology was published later and is available online as an Evaluation Toolkit. One example of additional guidance is the Decision Tree for Selecting the Evaluation Design.

The evaluation policy did outline several specific requirements for different types of evaluations, however, including:

- Each operating unit (OU) must conduct at least one performance evaluation of each large project it implements [12]
- Any project involving an untested hypothesis or demonstrating a new approach to development that hopes to be scaled up must be evaluated, ideally through an impact evaluation.

Several interview respondents noted that one of the most valuable aspects of USAID's evaluation policy was how clearly it outlined specific roles and responsibilities for various staff across the agency. The policy stated that every operating unit and technical / regional bureau must:

- Identify an evaluation point of contact who should be given an appropriate amount of time to devote to evaluation and training as needed.
- Develop a budget for evaluations, which should equal at least 3% of the program budget for the operating unit, on average.
- Prepare annual inventories of evaluations completed, as well as evaluations planned for the coming year.

In addition, the responsibilities of the new Learning, Evaluation, and Research office were clearly laid out:

- Ensuring human resources are sufficient across the agency.
- Investing in evaluation training and tools.
- Leading on impact evaluations, ex-post evaluations, and meta-evaluations on priority topics for the Agency.
- Prepare annual report on evaluation practices, changes, or updates.

Numerous interview and survey respondents said there have been large improvements to USAID's evaluation practices since the evaluation policy was adopted. Every single survey respondent from USAID agreed that since 2011 there is "more clarity about the roles and responsibilities of staff regarding evaluations." The vast majority also agreed that there is "more leadership support for devoting time and energy to evaluations"; "more staff who are appropriately trained and skilled at designing, managing, and assessing evaluations"; "more budget for evaluations"' and "more completion of high quality performance evaluations." There were mixed opinions on whether more high-quality impact evaluations have been completed since the adoption of the evaluation policy.

These responses are consistent with literature published by USAID, which states that since the adoption of the evaluation policy 1,600 USAID staff have received training in evaluations and the number of evaluations completed has increased from an average of 134 a year from 2006–2010 to an average of 236 a year from 2011–2015. [13]

Independent evaluators also agreed that since 2011 there are "more staff who are appropriately trained and skilled at designing, managing, and assessing evaluations" at USAID. Many of them spoke of dramatic increases in the sophistication of evaluation scopes of work (SOWs) coming out of USAID since 2011. "It used to be the 2-2-2- model," one evaluator said, "You send two consultants to look at two projects for two weeks, and then they write highly subjective impressions of what they saw. Now there is a real interest in evaluation methodologies and evidence in performance evaluations. Importantly, USAID is more likely to include the funds and time needed for actual data collection now." Another agreed, "There seems to be more understanding of what is needed to conduct a rigorous evaluation. The estimated levels of effort are getting better. They're still not adequate often but they're improving."

An OIG Audit of USAID's Evaluation Policy in 2015 found that USAID operating units had improved evaluation practices since the adoption of the policy. It found that operating units "generally complied with most of the . . . evaluation policy provisions. For example, they had established points of contact, developed mission orders for evaluation, and conducted evaluations when programs were innovative; in addition, PPL/LER had hired external evaluators to assess the quality of evaluation reports and how their findings were used."

However, the audit did find some areas where the agency was not in compliance with the policy, including the fact only 23% of operating units were dedicating the recommended 3% of their budget to evaluations, only 58% of missions had followed the requirement of evaluating every large project, and the majority of 'impact evaluations' reviewed by the OIG did not meet the definition of 'impact evaluation' – namely they did not determine a counterfactual.

After reviewing IG and GAO audits, as well as studies on evaluation quality and evaluation use, USAID updated its evaluation policy in October 2016. The updated policy modified requirements for when evaluations take place, strengthened requirements around evaluation use and dissemination, and encouraged evaluations of higher level outcomes.

# B. State Department

Almost a year after USAID adopted its evaluation policy, the State Department adopted a similar one in 2012. It required State bureaus to complete two evaluations between 2012 and 2014 and to evaluate all large projects and programs at least once in their lifetime or every five-years. The policy also encouraged 3%-5% of program resources to be identified for evaluation purposes, although it noted that evaluation funding needs would vary based on the size of programs.

According to a 2015 OIG audit of the 2012 policy over 40% of bureaus did not conduct the required number of evaluations, however, in part because "evaluations were a new concept to most bureaus, and most bureaus did not yet have the staff or funding to complete evaluations."[14] Furthermore, 45% of the completed 'evaluations' reviewed by the OIG did not meet the required elements of an evaluation, as laid out by the evaluation policy. Instead, they were organizational assessments or monitoring reports, incorrectly labeled as evaluations.

The OIG audit, survey respondents, and interview respondents from the State Department noted that insufficient staff capacity was a significant barrier to compliance with the initial evaluation policy. "The State Department doesn't have program officers. This means that evaluations are just another task or responsibility tacked on to someone's full time job," said one US government employee. Indeed, the OIG audit found that over half of Bureau Evaluation Coordinators did not perform evaluation duties full time and over a quarter of bureaus said they "had no trained, experienced evaluator on staff." In addition, the two bureaus in charge of overseeing evaluations – the foreign assistance office and the budget bureau – were themselves understaffed and could not meet the demand for guidance on evaluations from other bureaus and offices.[15]

Outside of several bureaus, the State Department does not have a long history of designing, implementing, and evaluating foreign assistance programs. One respondent pointed out that the lack of career program officers means that staff may work briefly in strategic planning, performance management, or evaluation, but they are unlikely to continue using and developing that skill set over the course of their careers. This contributes to a culture where evaluations are often not prioritized and the capacity to manage them is limited.

"There is not a culture of evaluation at State,"

"There is not a culture of evaluation at State," was a statement echoed by several respondents both internal and external to the State Department. This perception was paired with many related concerns, including "there is not a culture of transparency", "information is seen as power and not shared widely", "using data in decision making hasn't really gotten (to the State Department) yet", and "there are fears that evaluations might reveal sensitive information or show poor results and programs might lose funding." One respondent described widespread resistance to the Evaluation Policy, which resulted in taking a full year to get it adopted, and called its implementation a "case study in change management."

Over the past five years, State has taken several steps to address these challenges. It has offered trainings on "Managing Evaluations" and "Evaluation Design and Collection Methods" - which over 700 staff have completed - and it formed an Evaluation Community of Practice. It also began focusing more sharply on bureaus and areas where compliance with the evaluation policy was low (due to either resistance or technical difficulties incorporating evaluations into their work.)

One of these areas, highlighted in a recent Congressional Research Service report, was military and security assistance. The CRS report states that "Military and security assistance programs under State Department authority have gone largely unevaluated. The strategic and diplomatic sensitivities of this type of aid present significant challenges for evaluators. Past efforts by State to contract independent evaluators for these programs were reportedly unsuccessful, with the unprecedented nature of the work creating high levels of uncertainty and perceived risk among potential bidders."

"State is currently leading an interagency effort to strengthen performance management in security sector assistance."

In response, State staff note they are working to address these issues. State is currently leading an interagency effort to strengthen performance management in security sector assistance, including working on planning, budget development, program design, and performance monitoring. In November 2016, the working group developed a Performance Management Framework, which serves as an agreement on common principle and best practices for planning and performance management in the security sector. The State Department also worked intensively with the Department of Defense in the drafting and adoption of DOD's new evaluation policy.

In addition to challenges related to evaluating security sector assistance, the State Department also faces unique challenges related to evaluating diplomatic programs (in contrast to development programs). In a discussion about impact evaluations, one respondent highlighted some of these difficulties: "The desired outcomes of some State programs are diplomatic or political. These outcomes cannot be measured through a traditional development lens. The focus should be less on . . . doing more impact evaluations, and instead focus on building the capacity for staff to make informed decisions on what types of evaluations are most appropriate for the specific context."

In January 2015, State revised its evaluation policy based on the lessons learned in the first two years of implementing the initial policy. The new policy requires only one evaluation per bureau per year and does not require any evaluations at the post level. It also removed any guidance on funding for evaluations and now states that "the international standard of 3 – 5% of program costs is unrealistic." State finds that most bureaus actually spend closer to 1 to 2% of program funds on evaluation.

In its review of the initial policy, State found that many challenges in implementation were related to insufficient program planning and design for monitoring and evaluation. In response, State created a program design and performance management toolkit for bureaus to use when managing foreign assistance programs. In fiscal year 2016, State also created a dissemination policy for completed evaluations, including guidance for both internal and external distribution, which required that non-sensitive evaluations be posted online within 90 days of completion.

# C. The President's Emergency Plan for AIDS Relief (PEPFAR)

In 2003, the Bush Administration created PEPFAR to address the global HIV/AIDS epidemic through prevention, treatment, and care. According to PEPFAR's website, it is "the largest (effort) by any nation to combat a single disease internationally." Unlike USAID, the State Department, and MCC, PEPFAR is not an independent agency but is an initiative implemented by multiple USG agencies, including the State Department, USAID, the Department of Health and Human Services, the Department of Defense, the Department of Commerce, the Department of Labor, and the Peace Corps. As such, it does not have its own evaluation policy, but it is subject to the evaluation policies of these implementing agencies.

Through the implementing agencies, hundreds of evaluations are conducted on PEPFAR projects each year. In 2012, the Government Accountability Office (GAO) reviewed a random sample of PEPFAR-related evaluations and found that although evaluations managed by the Department of State's Office of the US Global AIDS Coordinator (OGAC) generally adhered to common evaluation standards – including ensuring fully supported findings, conclusions, and recommendations – the majority of evaluations from PEPFAR country or regional teams included findings, conclusions, and recommendations that were either only "partially supported" or "not supported." The GAO also found that "not all evaluation reports are available online, limiting their accessibility to the public and their usefulness for PEPFAR decision makers, program managers, and other stakeholders." GAO therefore recommended that OGAC work with other implementing agencies to improve adherence to common evaluation standards and increase online accessibility of evaluation results.

In 2013, the Institute of Medicine (IOM) conducted a systematic evaluation of PEPFAR's overall performance and its effects on health, as required in PEPFAR's reauthorizing legislation. While the evaluation covered many aspects of PEPFAR's programming, in regards to PEPFAR's evaluation practices it recommended: "To better document PEPFAR's progress and effectiveness, OGAC should refine its program M&E strategy to streamline reporting and to strategically coordinate a complementary portfolio of evaluation activities to assess outcomes and effects that are not captured well by program monitoring indicators. Both monitoring and evaluation should be specifically matched to clearly articulated data sources, methods, and uses at each level of PEPFAR's implementation and oversight."

In response to the GAO and IOM findings, PEPFAR released its "Evaluation Standards and Practices" in 2014, which laid out evaluation practices that all PEPFAR implementing partners should use. The document acknowledged that many of its standards and practices were already enshrined in agencies' evaluation policies; however, it set out to ensure they were consistently implemented across all PEPFAR partner agencies and organizations. It also began publishing links to all of its evaluations, in the form of downloadable spreadsheets listing all evaluations conducted in the year.

In September 2015, PEPFAR released a second version of its "Evaluation Standards and Practices", which includes additional guidance for operational issues associated with the planning, implementation, reporting, and review of evaluations. The new version also included more information on roles and responsibilities, tools and templates for evaluations, and required elements of planning and reporting on PEPFAR-funded evaluations.

Along with improving guidance on evaluations for its implementing partners, PEPFAR has also made great strides in collecting, analyzing, sharing, and using data that targets people at greatest risk in geographic areas with the highest HIV/AIDS burden. In 2015, PEPFAR stood up the Population-Based HIV Impact Assessments (PHIAs), which are household-based, general population surveys with HIV biomarkers, including measurements of HIV incidence, prevalence, viral suppression, and key behavioral indicators. These surveys provide critical information to assess the impact of programs on the HIV epidemic and identify programmatic areas that need improvement, as well as populations that need to be better reached. While these surveys are not evaluations in and of themselves, they serve a similar role for PEPFAR, allowing the initiative to understand epidemic changes and the impact of programs.

PEPFAR is currently collecting quarterly HIV epidemic and programmatic data down to the site level, which are then reviewed by country teams along with Ministries of Health, civil society, and headquarters to identify where mid-course corrections are needed. Performance and impact evaluations are then used as complimentary tools to receive additional qualitative information and explore case studies. "The data tell us if something is working or not," said one staff member, "And the evaluations help us to understand why our approach is working or not."

# D. MCC

As highlighted above, because of the parameters around which it was created under the Bush Administration, MCC already had many strong evaluation practices in place in 2009. Unlike USAID or State, the revised evaluation policy in 2012 was not meant to rebuild neglected (or non-existent) evaluation systems, practices, or capacity. Instead, it was meant to codify and standardize many of the evaluation practices already in place. According to the Congressional Research Service, "Since its inception, MCC policy has required that every project in a compact be evaluated by independent evaluators, using pre-intervention baseline data. MCC has also put a stronger emphasis on impact evaluation than State and USAID."

The agency's first set of independent impact evaluations – five studies focused on farmers' training activities – was wrapping up around the time the evaluation policy was updated in 2012. According to a World Bank study, only three impact evaluations using experimental designs in farmer training had been completed anywhere in the world from 2001 to 2011.[16]

Demonstrating its commitment to transparency, MCC published press statements, issues briefs, academic articles, and reports on the evaluations as a body of evidence in farmers' training. In addition to presenting the findings publicly, the agency used their release as an opportunity to educate stakeholders. They discussed what they learned about impact evaluations, what they learned about agricultural training, and what they learned about measuring household incomes.

Civil society advocates praised MCC for "taking a tremendous step forward in improving transparency and effectiveness within U.S. development programs." One interview respondent from civil society said "It was amazing. They put up everything. We were just awash in information, and we were learning so much."

There were initial concerns that important stakeholders (including Congress) would compare the positive findings of other agencies' performance evaluations to the more mixed findings of an MCC impact evaluation and conclude that the other agency's project must have been more successful. MCC's five impact evaluations found that MCC was "very successful in meeting or exceeding . . . output and outcome targets."[17] But, as impact evaluations, they went on to test whether the successful implementation of the projects resulted in changes to farmers' farm or household incomes. In three countries, farm incomes increased. However, none of the evaluations was able to detect changes in household income. That is to say, even though the projects were implemented well (as measured by exceeding output and outcome targets) they did not appear to lead to the desired impact (increased household incomes.)

"In addition to presenting the findings publicly, the agency used their release as an opportunity to educate stakeholders."

Thus, part of the learning for MCC was based around what hadn't worked about the impact evaluations. There were three primary issues that came up in implementing (and then interpreting results from) the impact evaluations – involving selection bias, adequate sample size, and homing in on the right evaluation questions to ask. "We weren't asking all the right questions," said one former USG employee. "We had all this great statistical data but not nearly enough qualitative information to determine why incomes didn't increase. It really made us re-think the importance of well-designed performance evaluations."

Indeed, despite the widely positive reception of MCC's release of its first five impact evaluations, the 2012 revision of MCC's evaluation policy downgraded

the role of impact evaluations. Instead of being an "integral part" of MCC's focus on results, the evaluation policy now recommended impact evaluations "when their costs are warranted."

Since 2012, MCC has continued to publish evaluations, summaries of findings, and underlying datasets online. However, it has not been as proactive in ensuring evaluations findings are aggregated across sectors, shared broadly with internal or external stakeholders, or used as a 'public good' to proactively inform policy-makers or practitioners. Since the initial launch of the first five impact evaluations, MCC has published 14 additional completed impact evaluations and 38 performance evaluations with little fanfare. These include five additional impact evaluations related to agriculture and irrigation, three in education, and three in transport. There are many more planned for publication in the coming years. To date, there is minimal evidence that MCC intends to maintain the high standard it set for itself in 2012 in aggregating findings across these sectors, publishing lessons learned or issue briefs, or hosting (or joining) events meant to publicize these findings are widely as possible.

# E. The Foreign Aid Transparency and Accountability Act (FATAA)

In July 2016, Congress passed the Foreign Aid Transparency and Accountability Act (FATAA), a bill originally introduced in 2011. Much like MCC's revised evaluation policy, FATAA did not create many new requirements or practices for most agencies, but it did codify many of the reforms the agencies had adopted recently. It also strengthened the evaluation requirements for many agencies that have smaller foreign assistance portfolios. "FATAA created a legislative mandate for evaluations," said one Congressional staffer, "This means that even as administrations and priorities change, the progress made in monitoring and evaluating foreign aid outcomes will be preserved." According to MFAN's website, "The bill institutionalizes the important gains that have been made on increasing aid transparency and improving monitoring and evaluation practices, and will help to ensure that this progress is built upon (by the) new Administration and Congress."

USAID, MCC, and State have all updated their evaluation policies explicitly to address FATAA requirements. Since FATAA mostly codified existing practices at USAID and MCC, the revisions to both their policies were relatively minor and can be found on the USAID website and the MCC website, respectively.

However, FATAA played a much more significant role in shaping the 2015 revision of State Department's Evaluation policy. According to one respondent who worked on the revisions, the likely passage of FATAA was "the single most important lever" in ensuring a permanent evaluation policy at the State Department.[18] "I can't say enough how important (FATAA) was to the revised evaluation policy. Without FATAA there would be no (permanent) evaluation policy and without the evaluation policy, there would be no evaluations at State," the respondent said, "(FATAA made it clear that) evaluations are not going away. We have to do this."

Other respondents felt that this was overstating the impact of FATAA, as there were already staff inside State working to strengthen and institutionalize evaluation practices. They did feel that FATAA was helpful to strengthening evaluation at State, however, due to its emphasis on the overall performance management cycle, including planning for programs, monitoring them for performance, and conducting evaluations. **"A major contribution of the FATAA legislation is that it placed evaluation within the continuum of performance management,"** said one respondent, "Without adequate planning, program design, and monitoring, it is difficult to get useful or reliable data from evaluations. FATAA clarifies that planning and design are as integral to learning and accountability as evaluation."

According to the legislation, OMB has until January of 2018 to set forth "guidelines - according to best practices of monitoring and evaluation studies and analyses - for the establishment of measurable goals, performance metrics, and monitoring and evaluation plans that can be applied with reasonable consistency to covered United States foreign assistance." The guidelines laid out by OMB could play a significant role in ensuring the recent progress in evaluation policies and practices is not only preserved but expanded upon.

# 03 Quality of Evaluations

In discussions about the state of evaluations at USAID and State prior to 2011, there are frequent references to the low quality of the evaluations. In contrast, MCC promised from its inception high quality evaluations. But what do people mean when they refer to the quality of evaluation? What metrics are used to assess the quality of evaluations? How can stakeholders determine the relative quality of evaluations at an agency, either over time or in comparison to other agencies? This section addresses all these questions using existing literature, survey findings, and interview responses.

There have been a number of recent attempts to assess the quality of evaluations in USG foreign assistance using set criteria and quantitative data. One of the first was a meta-evaluation on evaluation quality commissioned by USAID in 2013. This meta-evaluation used 37 criteria to assess the quality of 340 randomly selected USAID evaluations between 2009 and 2012 (before and after the adoption of the evaluation policy.) Almost all of the evaluations were performance evaluations (97%), which was consistent with USAID's portfolio of evaluations at the time. The findings are presented in Annex 3 with the 2009 data, and the 2012 data is listed beside the criteria.

The evaluator, MSI, pointed out that 68% of criteria showed an improvement from 2009 to 2012 and noted that the average composite "quality score" improved from 5.56 to 6.69 (on a scale of zero to ten) during the same time period.[19] However, only a quarter of the criteria showed an improvement of over 10 percentage points. The rest either showed small increases (which often fluctuated from year to year) or actually declined.

# Of those that did show notable improvements, three stand out as particularly important:

The evaluator, MSI, pointed out that 68% of criteria showed an improvement from 2009 to 2012 and noted that the average composite "quality score" improved from 5.56 to 6.69 (on a scale of zero to ten) during the same time period.[19] However, only a quarter of the criteria showed an improvement of over 10 percentage points. The rest either showed small increases (which often fluctuated from year to year) or actually declined.

- **The percent of USAID evaluations where management's purpose for undertaking the evaluation is explicitly described increased from 70% to 81%.** A common issue referenced in relation to the utility of evaluations is that they are conducted as a "box-ticking" exercise – completed only for compliance rather than due to interest ensuring accountability or learning. The relatively high percent of evaluations where the management's purpose for the evaluation is included – combined with the fact it appears to be on an upward trajectory – is encouraging.
- **The percent of USAID evaluations where study limitations were included increased from 38% to 64%.** One of the best tools for assessing the rigor and appropriateness of an evaluation's methodology is a transparent description of the evaluation's limitations. These limitations may include issues such as a lack of baseline or comparator data, or logistical barriers to conducting random sampling (such as conflict zones or impassable roads.) The notable increase in evaluations that included limitations suggest evaluators, donors, and other stakeholders were paying more attention to methods by 2012.
- **The percent of USAID evaluations where the recommendations are specific about what is to be done increased from 58% to 77%.** As explored more in the next section, there are numerous barriers to using evaluations to modify ongoing projects or inform future projects. One of the barriers is that evaluations do not always include specific, action-oriented recommendations that donors or implementers can choose to act upon (or not.) The increase in specific, action-oriented recommendations bodes well for an increase in evaluation utilization.

# Less encouraging are findings related to the following:

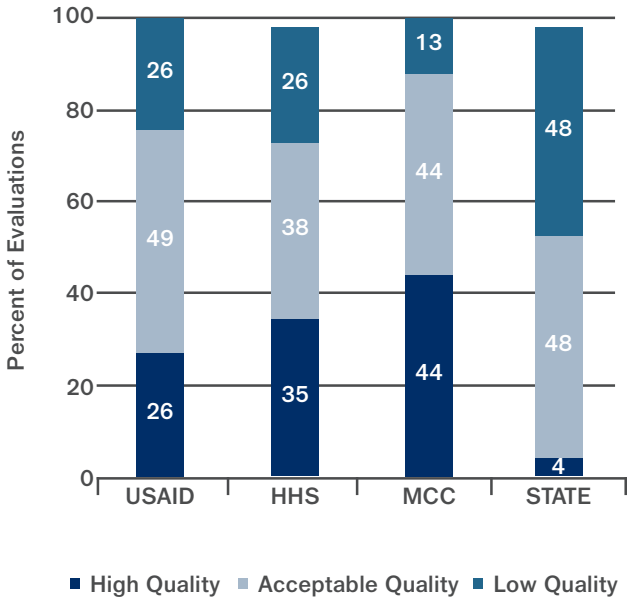
- **Fewer than 20% of USAID evaluations included an evaluation specialist: an individual, often holding a Ph.D., who is responsible for the design and implementation of research and reporting in order to demonstrate results of projects.** This is especially concerning because MSI finds that the average composite "quality score" is appreciably higher when an evaluation specialist is included (6.68 versus 5.56 on a scale of zero to ten).
- **Only about a third of USAID evaluations include local team members – both in 2009 and 2012.** In theory, governments, civil society organizations, and citizens of developing countries should be among the primary beneficiaries of evaluations. They have the most to gain from projects becoming more cost-effective and efficient in meeting their objectives. In practice, however, they are rarely included in evaluation teams and rarely included in discussions on how to learn or adapt from evaluation findings.
- **Despite requirements in USAID's evaluation policy, fewer than half of USAID evaluations discuss differences in project access or benefits based on gender.** Less than a quarter disaggregate evaluation findings based on gender. Both USAID's evaluation policy and many development stakeholders have pushed the importance of documenting and analyzing the impact projects have on intended female beneficiaries, specifically, who often face greater levels of poverty and more structural barriers to accessing the intended benefits of development projects.
- **Only 10% of USAID evaluations explore alternative possible causes of the results or outcomes it documents.** As previously noted, 97% of assessed evaluations were performance evaluations and therefore not designed to determine a counterfactual or prove attribution. However, it is still possible in performance evaluations – and, in fact, prudent – to ensure stakeholders understand that alternative factors outside of the project itself (such as projects from other donors, changes in government policies, weather patterns, etc.) may have influenced the documented results or outcomes.
- **Less than half of USAID evaluations include who is responsible for acting on each recommendation.** While there has been a notable uptick in the number of specific, action-oriented recommendations in USAID evaluations, it is not always clear who is responsible for acting on these recommendations. To the extent that evaluations are clear about who is responsible for carrying out recommendations (donors, implementing partners, partner governments, evaluators) it makes it easier to hold the responsible party accountable and more likely that recommendations will be acted on.

Similar meta-evaluations over time have not been conducted for MCC, State Department, or PEPFAR. However, USAID has recently committed to conducting meta-evaluations every two years, and in its fiscal year 2017 committee report the Senate Appropriations Committee required the State Department to conduct a review of the quality and utilization of the program evaluations of its 2015 programs. A recent GAO report compares the quality of evaluations in all these agencies (although not over time.)[20] It uses 8 criteria to assess the quality of 136 evaluations from USAID, MCC, State, and HHS completed in FY15.[21] The findings are presented below.

| | USAID | MCC | State | HHS |
|---|---|---|---|---|
| **Number of evaluations assessed by the GAO** | 63 | 16 | 23 | 34 |
| Study questions align with the key stated goal(s) of the intervention | 96% | 94% | 96% | 85% |
| Study questions align with the key stated goal(s) of the intervention | 96% | 94% | 96% | 85% |
| The chosen indicators/measures are appropriate for the study objectives. | 83% | 88% | 48% | 97% |
| The evaluation design is appropriate given the study questions. | 80% | 88% | 57% | 79% |
| The target population and sampling for the evaluation are appropriate for the study questions.goal(s) of the intervention | 59% | 63% | 43% | 62% |
| The data collection is appropriate for the study questions. | 63% | 69% | 35% | 76% |
| The data analysis appears appropriate to the task. | 63% | 69% | 48% | 74% |
| Conclusions are supported by the available evidence. | 73% | 75% | 61% | 65% |
| Recommendations and lessons learned are supported by the available evidence. | 65% | 79% | 86% | 80% |

Overall, GAO found that most evaluations were either "high quality" or "acceptable quality" based on the above criteria. However, there was variance across agencies. In general, the area where evaluations were least likely to meet the criteria was in the design and implementation of the evaluation's methodology: specifically, sampling, data collection, and analysis. Almost 40% of evaluations either did not describe how respondents were selected for surveys, interviews, and focus groups or used sampling methodologies with limitations. According to the GAO, "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."

Nearly 40% of evaluations also had limitations in, or provided insufficient information about, their data collection methods. These limitations included a lack of transparency about data collection instruments (such as survey or interview questions), a lack of baseline data, or a lack of targets to compare progress against. Similarly, about 40% of evaluations had limitations or lack of transparency about their methods of data analysis. These limitations included a lack of transparency about the types of statistical tests, sensitivity analyses, or robustness checks used by evaluators to analyze their data.




When asked to define a 'quality evaluation' for this assessment, several interview respondents discussed the importance of methods that ensured objectivity and reliability. "I want to know that if we sent another evaluation team to conduct the same evaluation using the same methodology, they would come back with the same findings," one former USG employee said. As highlighted earlier in the reports, a great deal of criticism about evaluations related to low levels of rigor in evaluation methodology. The "2-2-2- model" and the "go take a look-see" approach are scorned because they are not systematic and are highly subjective. The limitations highlighted in the GAO report suggest that these issues have not been fully resolved.

"I want to know that if we sent another evaluation team to conduct the same evaluation using the same methodology, they would come back with the same findings,"

These issues are highlighted with even more concern in a recent paper from CGD. In their paper entitled "How Well Are Aid Agencies Evaluating Programs? An Assessment of the Quality of Global Health Evaluations", the authors assess 37 performance and impact evaluations of global health programs carried out by five agencies, only two of which (USAID and PEPFAR) are covered in this assessment, based on criteria related to relevance, validity, and reliability.

They found that most evaluations were based on questions relevant to the health program under evaluation. However, only 38% of impact evaluations and no performance evaluations used the most relevant data to answer those questions.[22] If only some of these criteria were met evaluations were given a "medium" score: 50% of impact evaluations and 33% of performance evaluations were assessed as "medium" on data relevance. **Over two-thirds of performance evaluation did not meet any of the data relevance criteria.**

In terms of reliability, only 38% of impact evaluations and 10% of performance evaluations followed accepted social science methods for sampling, including using random sampling that was purposeful, adequately justified, and included heterogenous populations. An additional 13% of impact evaluations and 10% of performance evaluations received a "medium" score; they used convenience or purposeful sampling but only partially met the criteria of justifying their sampling

approach and seeking heterogeneous populations. **The remaining 50% of impact evaluations and 81% of performance evaluations were either not transparent about their sampling methodologies or did not meet the criteria.**

The authors also assessed whether evaluations took a systematic analytical approach and considered potential confounding variables. Only 31% of impact evaluations and 5% of performance evaluations had high analytical validity and reliability. In order to be assessed as having "high validity and reliability", an evaluation needed to fully describe its randomization method (if relevant) and consider all appropriate covariates that may influence its findings (whether randomized or not.) If covariates are only partially considered, an evaluation was given a score of "medium": 50% of impact evaluations and 24% of performance evaluations were assessed as having "medium" analytical validity and reliability. **71% of performance evaluations did not consider covariates that may have influenced their findings.**

It is important to note the CGD paper evaluated five aid agencies. However, its findings appear generally consistent with the MSI and GAO papers, although the exact criteria and standards used to assess the evaluations vary across all three studies.

For example:

- The GAO finds that about 40% of evaluations fail to provide evidence of adequate data collection to answer the evaluations' questions; CGD finds that 44% of health evaluations did not collect relevant data.
- The GAO finds that about 40% of evaluations fail to meet adequate sampling techniques; CGD finds that about 68% fail to do so.
- As for considering covariates or alternative possible causes for evaluations findings, MSI stated that 90% of USAID evaluations did not consider alternative possible causes of the results or outcomes they document; CGD states that 49% of health evaluations did not adequately consider covariates.

While this assessment has been clear that there is an important role for both performance and impact evaluations in ensuring accountability and learning, the CGD paper does suggest that – in practice – many impact evaluations of health interventions are currently designed and implemented with a higher level of relevance, validity, and reliability than many performance evaluations of health interventions. According to CGD's assessment of health interventions, "Impact evaluations are achieving better methodological rigor even though deficiencies persist. By contrast, methodological rigor in performance evaluations is lacking. Given that there are so many performance evaluations conducted, it is essential to improve the quality of these studies by encouraging less biased approaches to collecting information and greater rigor and replicability in the methods used to derive conclusions."[23]

**An evaluation's quality may be jeopardized if the evaluator has a conflict of interest.** When an evaluator or an evaluation organization has a conflict of interest, it can lead to less objective, rigorous, or useful evaluations. This was a concern voiced by several interview respondents. "Since a number of NGOs and consulting firms do both implementation and evaluation, it is not uncommon for evaluators to be assessing their competitors' programs." Another added, "A number of implementers are buying up evaluation firms, so this is only likely to become more of an issue."

In order to address conflict of interest, some donors ask evaluators to complete conflict of interest forms. However, this is still quite rare. According to MSI, only 12% of USAID evaluations included conflict of interest forms in 2012. By 2016, the GAO finds the number to be closer to 70% for HHS, 40% for USAID, 30% for State, and 0% for MCC.[24] The CGD study found no conflict

of interest forms in any of the health evaluations. An absence of conflict of interest forms does not necessarily signal a conflict of interest is likely. In CGD's assessment, the authors were not able to identify clear conflicts of interest for 88% of the evaluations they examined. However, the authors did identify potential conflicts of interest in 12% of evaluations, which signals this is an issue worth considering further.

**While independent evaluations are important, this does not mean there needs to be a firewall between evaluators and implementers or donors.** The logic for prioritizing independence is clear: by ensuring evaluators have no stake in the project they are evaluating, it aims to minimize positive bias in the evaluation. This is generally agreed to be an important step in ensuring more objective evaluations. However, stakeholders do not agree about what level of independence is best. While most agree that people should not be evaluating their own projects - or their direct competitors' projects- many feel there is a lot to lose by enforcing absolute independence.

First of all, in many impact evaluations, evaluators must work in conjunction with implementers through the entire life cycle of the project – from project design to completion – to ensure the success of the evaluation. (Note the example of farmers' training in Honduras above to see what happens if there is insufficient cooperation between implementers and evaluators.) Second, many people believe that the more involved donors and implementers are in the evaluation process, the more they are likely to utilize evaluation findings for learning and accountability. This suggests that while donors should take great measures to ensure there is no conflict of interest for the evaluators, it may not be beneficial to ensure a complete firewall between evaluators and implementers.

# 04 Utilization of Evaluations

Ultimately, **evaluations are only worthwhile if they are actually used to improve development effectiveness.** Evaluations need to be utilized by various stakeholders to inform decisions if they are to make any difference. When asked what types of conditions lead to evaluations being used, interview respondents' answers generally fell into three categories:

- Evaluations must be built around rigorous and objective methodologies and deliver credible findings, lessons learned, and recommendations.
- Evaluations must be accessible - in both a literal and figurative sense - to a wide array of stakeholders, including program designers, program implementers, policy-makers, and citizens in both the donor and beneficiary countries.
- There must be opportunities, incentives, and systems in place to design and modify programs based on evaluation findings.

The current status of each of these conditions – and the barriers to achieving these conditions - is explored below.

- **Evaluations must be built around rigorous and objective methodologies and deliver credible findings, lessons learned, and recommendations.**

The previous sections suggest that since the adoption of evaluation policies at USAID and State, as well as the revision of the evaluation policy at MCC, there has been progress towards increasing the number of evaluations completed, as well as increasing the quality of some evaluations. However, there is still a lot of room for improvement when it comes to the rigor, objectivity, and credibility of many evaluations. This is important because many interview respondents said they are more likely to use evaluations if they believe the methodology is rigorous and the findings are credible.

**In order to produce quality evaluations, it is necessary to design projects that can be evaluated well.** This means clearly articulating the theory of change, causal logic, and desired outputs, outcomes, and impacts of the projects. It means using relevant indicators to track progress towards these targets. In addition, it means considering who the intended audience of the evaluation is and how it will be used from the beginning.

A number of interview respondents highlighted that some of the value of evaluations comes from the process and not just the end product. They pointed out that incorporating M&E systems and tools into program design tends to make the design itself better - it tends to ensure better articulated theories of change, more explicit discussion of causal logic, and more clear and quantifiable metrics of success. "Regardless of what is measured at the end; the fact that it is measured changes things," said one USG employee.

Another agreed, "The back and forth process (between) Mission staff and evaluators is where a lot of the actual learning occurs. Even if some of the negative findings are left out of the final report, they were often shared, discussed, and maybe even internalized by the Missions first."

One former USG respondent discussed the role leadership can play in ensuring the process of incorporating M&E systems into program design has on improving the quality of program design. The respondent discussed the fact that at the agency, leadership would ask detailed questions about evidence informing the project design, as well as robustness of M&E systems before approving the project. "Knowing they would have to answer questions for the Investment Committee Meeting made (the staff) more rigorous in their project design."

It may also be the case that causation runs in both directions: increasing the quality of evaluations raises levels of utilization, and similarly increasing the utilization of evaluations leads to donors and implementers demanding higher quality evaluations, as they become increasingly aware of what they need to make evaluation findings credible and useful. "There may be a lag time," one respondent said. "Many stakeholders are still becoming aware that there are more and more quality evaluations out there. It may take them a little while to realize how useful they can be, especially if they're used to lower quality evaluations that are minimally helpful."

**The evaluations must be accessible - in both a literal and figurative sense - to a wide array of stakeholders, including program designers, program implementers, policy-makers, and citizens in both the donor and beneficiary countries.**

In "Beyond Success Stories," the authors describe the state of evaluation utilization in 2009: "Evaluations do not contribute to community-wide knowledge. If "learning" takes place, it is largely confined to the immediate operational unit that commissioned the evaluation rather than contributed to a larger body of knowledge on effective policies and programs."

In 2016, USAID commissioned a meta-evaluation of evaluation utilization, which found that 93% of sampled evaluations at USAID had stimulated learning and 90% influenced decisions or actions.[25] While this sounds like a very different environment and situation from the one described in "Beyond Success Stories," a deeper dive into the data shows similar themes. The statistics above come

"still a long way to go in terms of using evaluations to test theories of change or fundamentally challenge development assumptions."

from surveys of 118 USAID staff answering questions about evaluations of which they were previously aware. It turns out most of them (72%) were aware of the evaluation because they had been involved with planning it. This suggests that the majority of evaluation utilization still takes place largely within the immediate operational unit that commissioned the evaluation.

This is not to dismiss the positive finding that USAID staff involved with planning evaluations do subsequently use those evaluations to inform future decisions or actions. Indeed, that is one of the primary purposes of evaluations. Furthermore, about 66% of survey respondents said post-evaluation meetings were called to decide how to respond to the evaluation's findings or recommendations, which suggests the evaluations were not just read and dismissed, but also used to inform decisions.

However, the survey findings suggested there was **still a long way to go in terms of using evaluations to test theories of change or fundamentally challenge development assumptions.** Of the 609 evaluations explored in the meta-evaluation, only three were impact evaluations and only 6% of survey respondents said the purpose of the evaluation was to "test assumptions or hypotheses." Even among the performance evaluations, 72% of survey respondents said the evaluations "confirmed what USAID already knew," and 66% said the projects being evaluated performed "as expected."

In addition, these data beg the question of whether other stakeholders use the evaluations. According to the meta-evaluation, USAID staff generally received full access to briefings and briefing materials, and the majority of implementing staff also had access. However, partner countries, beneficiaries, and other stakeholders had access less than a third of the time. While similar meta-evaluations have not been completed for MCC, State Department, or PEPFAR, the following section relies on surveys, interviews, and websites to explore accessibility across all four organizations.

**Although USAID, State Department, PEPFAR, and MCC all publish evaluations online, they are not always presented in a user-friendly format.** The question of access is an important one for evaluation utilization. As highlighted in the section on MCC's evaluation policy, the first step many donors take to ensure access is publishing their evaluations online. (Unfortunately, this is often the last step as well.). USAID publishes its evaluations on the Development Experience Clearinghouse (DEC.) The DEC contains five decades worth of documents capturing USAID's experiences with development, including but not limited to evaluations.

"USAID publishes its evaluations on the Development Experience Clearinghouse (DEC.)....numerous interview respondents complained about the lack of user-friendliness in the DEC."

While USAID deserves credit for publishing the vast majority of its evaluations online, numerous interview respondents complained about the lack of user-friendliness in the DEC. One civil society advocate said, "I was working on a paper for public-private partnership so I went to the DEC to look for evaluation findings that might inform my

paper. When I searched 'public-private partnership' 44,000 results appeared. Obviously, I was overwhelmed, but I went ahead and read through a couple. I quickly realized most of them weren't even evaluations, they were just documents that mentioned 'public-private partnerships' in passing."

Another agreed, "I try to use evaluations to inform my organization's policy agenda, but it's not easy. The DEC is a mess." "Even if you can find a relevant evaluation," one respondent said, "the variance in methodology and quality of evaluations at USAID means you can spend hours just trying to figure out what you're looking at."

As of the drafting of this report, the DEC does allow you to filter down to just evaluations, of which there are 11,514 posted. In order to filter further, one has to enter search terms. For example, a search for "education" results in 6,969 evaluations. A search for "girls' education" results in 154 evaluations. A search for "girls' literacy" results in 22 evaluations, which seems like an appropriate body of work for a preliminary literature review. This suggests it is possible to find relevant evaluations on the DEC for specific sectors, but one has to be relatively precise about what one is interested in reading about. Otherwise, the magnitude of documents on the DEC can be unwieldy.

At the State Department, on the other hand, evaluations weren't posted online until 2015. As of the drafting of this report, there are fewer than 60 evaluations posted. One former USG employee said, "Getting the evaluations online at all was a heavy lift. There was a lot of push back. Staff were worried they would reveal sensitive information, or show bad results, or just be embarrassingly poor quality." To address the concerns about revealing sensitive information, State requires that bureaus completing evaluations deemed sensitive provide a "summary of results" that can be published in lieu of the complete evaluation. A civil society advocate noted, "Even once State decided to publish evaluations online, timeliness was a serious issue. They wanted to publish evaluations on an annual basis. We had to push them hard to include the requirement that evaluations be published within 90 days." (Staff from the State Department note that the practice of publishing evaluations annually was tied to internal reporting requirements and IT systems capabilities, which have since been improved to facilitate more frequent posting.)

PEPFAR publishes its evaluations on its website via downloadable spreadsheets. These spreadsheets list all the evaluations conducted in the year, sortable by organizational unit, evaluation title, evaluation questions, implementing agency, and the link to the evaluation. According to the website, the spreadsheets are updated "intermittently" with 2016 evaluations being posted as of April 2017.

MCC publishes its evaluations on the Evaluation Catalogue. They can be filtered by country, topic, year, or inclusion of datasets. There are currently 119 evaluations posted, and many more are scheduled for the coming years (A list of MCC's planned evaluations can be found on their website.). Along with the completed evaluations, MCC also publishes the underlying datasets (when relevant) and Summaries of Findings (when available.)

Summaries of Findings are shorter documents (often around six pages) that summarize the project's context, program logic, monitoring data, evaluation questions, evaluation results, and lessons learned. It is not uncommon for evaluations to exceed 100 pages. They often include methodological and technical details, footnotes, and annexes that are vital to ensuring the evaluation's quality, but that make the evaluations less accessible to stakeholders who are not evaluators or M&E experts. While most evaluations include short executive summaries, the content of these executive summaries can vary dramatically based on the topics covered in the evaluations. MCC's Summaries of Findings, on the other hand, use a standard template to pull the same information from each evaluation. This makes it easier to know exactly what information is included, find specific topics of interest, and compare across evaluations.

"In order for evaluations to be useful, policy-makers need one-pagers focused specifically on learning and policy change."

**Even when stakeholders can find evaluations, they may still struggle to find the information they need from the evaluations.** MCC's Summaries of Findings are important because literal accessibility is only the first step towards transparency and evaluation utilization. Evaluations must also be accessible in a figurative sense. "It's a constant battle," said one interview respondent, "On the one hand you have the M&E folks and evaluators who are deeply concerned about maintaining absolute rigor and accuracy in the evaluation's findings. But on the other hand, you have the public affairs people who really just want three great bullet points." Another gave a specific example, "We had evaluators who insisted the executive summary had to be 25 pages long because cutting it down would result in 'losing some of the nuances.' But you lose all the nuance if no one ever reads it!"

One civil society advocate urged, "In order for evaluations to be useful, policy-makers need one-pagers focused specifically on learning and policy change. They shouldn't include anything on methodology or project descriptions . . . just one page of action-oriented recommendations that are displayed, organized, and shared in a proactive and intuitive way."

This should not be construed as backtracking on the importance of methodology emphasized in the previous section. "The evaluations must be valid and reliable," one senior manager said, "But once that has been determined by the experts, just give me the top line findings that I can actually act upon."

"It's like your bathroom plumbing," one civil society advocate said, "You need it to be well-designed and fully-functional. But that doesn't mean that you personally need to know how it works or that you want to see the inner workings."

While the need for short, action-oriented summaries was repeated across managers and policy-makers, it was emphasized most often in regards to Congress. "We don't need to know the details of the methodology or even the findings," one Congressional staffer said, "But we want to know that someone knows. We want to know that someone is reading these evaluations and learning and moving money accordingly."

**Many evaluations are planned and published on an ad hoc basis, rather than as part of a systematic learning agenda.** Several respondents discussed the need for evaluations of similar projects to be clustered and synthesized. This requires both proactive planning before evaluations are commissioned and systematic efforts to synthesize findings after evaluations are completed. "In order for individual evaluations to roll in to something bigger, there needs to be a learning agenda," one respondent said, "Agencies need to know what questions they are trying to answer and then cluster evaluations around those questions. MCC did this with their first batch of agricultural evaluations. USAID's Democracy and Governance team set an annual learning agenda. But most agencies or offices don't do this."

Recognizing this, USAID recently published a Landscape Analysis of Learning Agendas: USAID/Washington and Beyond. The report highlighted and analyzed six learning agendas being used at USAID, including looking at the motivations, strategies, benefits, and challenges of designing and implementing a learning agenda.

Another USG senior leader advocated for an even broader plan, "The agencies should sit down together and determine 'what are the 15 key questions in each sector that we need answered?' And then they should prioritize evaluations that answer those questions. Right now, we are spending a lot of money on evaluations but they're individual and ad hoc, based on whatever a specific office is interested in or whatever project we happen to be spending money on."

MCC attempted to cluster and synthesize its findings with the agricultural impact evaluations released in 2012. Since that time, however, there have not been similar publications around their release of impact evaluations or around specific topics.[26] USAID's Bureau for Economic Growth, Education, and Environment has released a Sectoral Syntheses of Evaluation Findings and the Bureau for Food Security has an annual publication on results, both of which Missions can read (rather than perusing dozens of individual evaluations.) However, this is not a standard practice throughout the agency.

**There must be opportunities, incentives, and systems in place to modify programs based on evaluation findings.**

Even when evaluations are both high-quality and accessible, they still may not be used by stakeholders if there are not opportunities, systems, and incentives in place to encourage their use. The first condition to encourage use is ensuring there are opportunities to use evaluation findings to inform resource allocation and project design. This begs the questions "who controls project funds?" and "who designs the projects?"

**When funds for programs are inflexible, it is difficult to adapt programming based on evaluation findings or recommendation.** The State Department, USAID, and MCC all get their funding from Congress in the form of an annual budget. The budget process is long, complicated, and full of uncertainty. It is also full of Congressional earmarks, Presidential initiatives, and funds set aside to respond to current crises (refugee migration, famine, Ebola, etc.) According to George Ingram at Brookings, "Presidential initiatives have their place as a way to bring along political allies and the American populace. It is also appropriate and constructive for Congress to weigh in on funding priorities. But it can be counterproductive to effective development when presidential initiatives and congressional earmarks dictate at the micro level and restrict flexibility in implementation."[27]


"Many USG interview respondents voiced frustration at earmarks, particularly in relation to USAID funding."

Many USG interview respondents voiced frustration at earmarks, particularly in relation to USAID funding. The frustration was primarily related to feeling micro-managed with inflexible budgets, which allow limited space for evidence-based policy-making within the agencies. However, several respondents also pointed out that despite demanding more and better evaluations (and evaluation use) Congress itself has not shown an appetite for using evaluation findings themselves to inform budgets. "Congress should be much more interested than they are," said one USG employee.

"Policies are often influenced by politics and political process; not evidence," said another respondent, "So is there a demand for evidence to inform your policies? Is there space for it?" Although he was referencing policy-makers more broadly, he did also speak specifically about a Congressional appetite for evaluation findings. "If Congress isn't going to use the evaluation findings themselves, can they strike a balance with the agencies where they promise more flexible budgets and fewer earmarks if and when the agencies show they have evidence about 'what works' and a plan to modify programs accordingly?"

The questions around flexible budgets and program design don't stop with Congress. Agency staff also noted that inflexible procurement regulations make it challenging to hire quality evaluators. A number of implementing partners – both NGOs and private-sector – commented on "prescriptive Requests for Proposals (RFPS)" from donors that do not allow implementers the space to cite previous evaluation findings or evidence in their proposals. "If you want to win the bid, you're not going to challenge the donor's project ideas, even if you've seen evidence that it hasn't worked in the past, or evidence that another approach might work better."

"There is a lot of risk-aversion and aversion to change. We already have the mandate to continue on the status quo – making changes is hard."

When asked whether donors are open to learning from evaluations, one implementer said, **"There is a lot of risk-aversion and aversion to change. We already have the mandate to continue on the status quo – making changes is hard.** There are sometime turf wars within agencies or staff don't agree about how to make changes. It's easier to just continue with the status quo."

On the other hand, some implementers felt citing evaluation findings or evidence in their proposals actually made them more competitive (although it was not clear from context if this was in reference to challenging an existing project design or using evidence to support the existing design.)

Beyond inflexible budgets and project design, there are a number of ways that the opportunity for using evaluation findings is constrained. These include the timing of evaluations, a lack of political will or leadership, and lack of staff time and capacity.

**When evaluations are not well-timed, their findings are less helpful to decision-makers.** The timing of evaluations can be complex for several reasons. There are both benefits and costs to conducting evaluations at every stage of a project. Evaluations

conducted during project implementation can be helpful in ensuring implementation is going as planned and allowing donors and implementers to make course-corrections as needed. However, they are unlikely to tell stakeholders much about outputs, outcomes, or impacts because the project is not yet complete. Evaluations conducted at the end of the project can produce more information on final outputs and some outcomes, as well as comprehensively assess project implementation. However, they still might not be able to capture all outcomes or impacts because benefits of many projects need several years to manifest or accrue. In addition, if the evaluation uncovers serious issues, it is too late to make changes to that project. Evaluations conducted several years after the completion of a project (ex-post evaluations) allow evaluators to assess whether the project has been sustained and whether the intended benefits did accrue over the years. However, it becomes increasingly more difficult to attribute benefits to a specific project as increasingly more exogenous variables come into play.

There are also timing issues related specifically to impact evaluations, which came up frequently in interviews about MCC. "It is challenging to balance impact evaluation timelines – which tend to take many years – with the need for more immediate feedback," one USG employee said. Others were blunter: "By the time you get the data, is it still relevant? Since the evaluations were designed (in the mid to late 2000s), MCC's model has changed, and the governments we're working with have changed." While the findings from long-term impact evaluations may still be helpful in terms of providing evidence about whether an intervention is effective, staff at MCC generally felt that the timing and duration of many of the first impact evaluations made them less helpful in terms of project design and management.

"It is almost always better to plan evaluations during project design, before implementation has begun."

One timing consideration is true for all types of evaluations, however: it is almost always better to plan evaluations during project design, before implementation has begun. For impact evaluations, it is a requirement, because projects must be designed and implemented in a way that allows for a beneficiary group and a comparable control group. For performance evaluations, starting early allows for the collection of baseline data. In both cases, designing the project and evaluation simultaneously encourages a more analytical approach to determining the theory of change, the causal logic, and the indicators and targets for outputs, outcomes, and impacts. It also encourages collaboration between project designers and M&E specialists, which is an asset further explored later in the report.

**When leadership does not prioritize learning and accountability through evaluations, staff are unlikely to do so either.** Leadership serves as a major influence for evaluation utilization within an agency, office, or team. "The M&E person is not the most senior person on the team," said one independent evaluator, "You go to them when you want a chart; not when you want to make a decision. When you want decisions, you need to go to the top."

In the early days of the Obama administration, many leadership positions were filled

with champions of evaluations and evidence-based decision making. At USAID, Administrator Shah established the Office of Learning, Evaluation, and Research, brought in Ruth Levine to help draft the evaluation policy, brought in Michael Kremer (one of the pioneers of impact evaluations of development projects) to help USAID better pilot, test, and scale innovative approaches to development, hired more M&E staff and fellows, and set specific targets for increasing the number of evaluations commissioned at USAID. At MCC, CEO Yohannes brought in Sheila Herrling – the Director of Re-thinking Foreign Assistance at CGD – to serve as the Vice President of Policy and Evaluation and subsequently oversee the revision of MCC's evaluation policy and shepherd the publications related to the first five impact evaluations.

"Ultimately, it is leadership that is going to demand evidence-based project design (or not.)"

**Ultimately, it is leadership that is going to demand evidence-based project design (or not.)** When leadership asks questions about evaluation findings, or the evidence informing a project's design, or insists that evaluation recommendations be acted upon, it sends a message that evaluations matter.

**The lack of staff time and bandwidth is the single biggest constraint identified by survey and interview respondents. This is exacerbated by a lack of clear roles and responsibilities related to utilizing evaluation findings.** When asked about the primary barriers to evaluation utilization, one USG employee said, "Time, time, and time. There are trade-offs between 'learning' and 'doing.' And even though we conceptually understand that learning up front will make implementation easier later on, it is exceptionally difficult to find staff who have both the skill set and the time to think about how we could best share or apply evaluation findings."

The lack of time for staff to prioritize reading, sharing, and acting upon evaluations is exacerbated by the fact it's not clear who should be taking the lead on these tasks. It often falls to M&E staff, because they are the primary lead on many evaluations. However, in many agencies and organizations, M&E staff are not involved in making strategic policy, budget decisions, or designing subsequent programs. "There is a silo between M&E people and operations people," said one USG employee. **"A lot of learning is stored exclusively in one person's head and only gets shared or used on an ad hoc basis."** Conversely, program or operations staff are often the people designing projects, but it is rare for evaluation review to be a significant part of their portfolios. Communications or public affairs staff often want to message positive findings from evaluations, but they usually want only a few bullet points highlighting successes.

In most agencies and organizations, utilizing evaluations falls primarily to those who commissioned them, which is often M&E staff. But M&E staff are also in charge of all monitoring activities, as well as overseeing the implementation of evaluations. They are often overwhelmed by a number of competing requests and priorities, many of which are requirements forced on the agency or organization by external stakeholders. One M&E staff in an implementing partner organization said, "We are collecting 150

indicators for (our donor.) How would we even begin to analyze 150 indicators? Are they analyzing this data in some way? It seems like there is sometimes an overproduction of data with little clear purpose of 'data for what?'" The need to better select, harmonize, and share M&E indicators and data came up several times as a way to reduce both the cost and time burden of data collection for evaluations.

There was also a range of views on the best role for agency staff to play in implementing evaluations. Many interviewees felt that including agency staff in conducting an evaluation was important for buy-in and likely to lead to better utilization of the final report. Some felt that it was best for DC staff to be involved, since they were likely to have more distance from the project and therefore be more objective during the evaluation. Others felt that it was better to have mission staff involved because their investment in the project made them more likely to internalize and utilize any lessons learned. While there was not consensus about the best staff to include, most interviewees agreed that the benefits of including agency staff outweighed the costs of potential positive bias.

**There are rarely incentives for prioritizing evaluations or their utilization.** "I sat on the promotion committee for decades," said one former USG employee, "No one was ever promoted because they commissioned a great evaluation or used it to improve subsequent program. Evaluations never even came up in the discussion." Current USG staff agreed, "I have so much on my plate that is urgent. Reading and internalizing evaluations is never urgent. Time for it isn't built into [project design] for it. If we're racing deadlines, we never had time for reflection, much less active learning."

"It's so much easier to generate flashy success stories than generate knowledge, especially if you believe you'll be punished for a poor result."

This may be related to the fact there is not much incentive for agencies or their leadership to prioritize evaluation utilization and learning. "It pays to be ignorant," said one civil society advocate, **"It's so much easier to generate flashy success stories than generate knowledge, especially if you believe you'll be punished for a poor result."**

Despite widespread rhetoric encouraging accountability and learning, there are very few stakeholders actively working to incentivize it. Congress passed FATAA but made it cost-neutral, which means there is no new funding to increase the amount of time staff spend internalizing, synthesizing, sharing, or acting upon evaluation findings. Congress also has not reduced the number of reporting requirements in these agencies, which usually falls on the same staff.

Civil society advocates often talk about the importance of learning from evaluations but rarely utilize the evaluations themselves to hold agencies accountable to meet the promised objectives or to change their projects in the future. Several interview respondents discussed the role external stakeholder such as Congress or civil society could play in holding agencies responsible for using evaluation findings: "Can you imagine if someone created a rolling list of Lessons Learned from these public, independent evaluations? Or required the agencies to do so themselves? I think it

would be eye-opening. I mean, if you didn't learn from the first 499 evaluations that 'interventions should have a clear program logic that link(s) activities to their desired outcomes,' what makes you think you'll have a break-through the 500th time its mentioned?"

**Along with a lack of staff time and incentives, very few agencies, offices, or organizations have systems in place to ensure the utilization of evaluations.** Although there is evidence that learning does occur, it tends to happen in silos and in an ad hoc manner. "To the extent that evaluation findings are shared, it's through brown bags or hallway encounters," said one USG employee.

Some agencies, offices, or organizations do have some systems in place to encourage evaluation utilization. More than half of the survey respondents from USAID's meta evaluation of evaluation utilization in 2016 said the "majority of evaluations' recommendations were adopted or otherwise acted upon." About 22% of the survey respondents said a written Action Plan was employed to help address recommendations from evaluations. According to the updated ADS 201, USAID offices are now required to develop action plans for all evaluations.

In addition, a number of platforms have been developed to help stakeholders share and access findings from evaluations or research in order to encourage learning. For example, USAID's Bureau for Food Security hosts Agrilinks, a knowledge sharing platform for capturing and disseminating new learning in agriculture and food security. Similar platforms include WLSME for sharing knowledge around women's leadership in small and medium enterprises, Microlinks for sharing good practices in inclusive market development, and Learning Lab to help development professionals collaborate, learn, and adapt through learning libraries, learning networks, and other resources.

In early 2017, MCC began publishing a newsletter on evaluation findings and evidence, which was circulated to all staff and shared externally. Some donors have sponsored evidence summits to bring stakeholders together for several days, specifically to share and learn from evidence in a sector. For example, in the summer of 2016, MCC partnered with the Government of El Salvador to host an evidence workshop, attended by 180 policymakers, practitioners, and researchers interested in sharing evaluation findings related to education and the investment climate in El Salvador. Brown bags are also used to encourage internal learning from evaluations, although some respondents felt like this type of venue only strengthened the impression that using evaluation findings is an optional, extra task for staff, rather than a mandated part of their portfolios.

# 05 Recommendations

A great deal of progress has been made in establishing better evaluation practices of USG foreign assistance within USAID, the State Department, and MCC. However, much work must still be done to improve the quality and use of evaluations. In addition, there is danger of the recent progress stalling or even being reversed if steps are not taken to prioritize evaluation practices and utilization.

In order to ensure further progress, the new administration, Congress, and agency leadership should adopt the following recommendations.

## Evaluation Policies

- The new administration should ensure agencies continue to implement evaluation policies by appointing leadership with a demonstrated commitment to using evaluations to inform evidence-based policy-making. Currently, the administration is proposing significant cuts to foreign assistance. One way to ensure that the State Department, USAID, PEPFAR, and MCC can maximize the impact of every US dollar is by encouraging accountability and learning through the utilization of evaluations. As of the drafting of this report, many leadership positions are still vacant at all of these agencies. The administration should ensure these positions are filled with people who understand the value evaluations can play in ensuring foreign assistance is as cost-effective and evidence-based as possible for the strongest outcome for aid recipients.
- Congress should provide rigorous oversight of the implementation of FATAA both as OMB writes the agency guidance for it and as the agencies carry out this guidance. Congress can ensure agencies maintain and advance the progress they've made on evaluation practices by enforcing FATAA. This includes ensuring strong guidance is provided for the implementation of FATAA and requiring agency leadership to give updates on the progress they are making in meeting the objectives laid out in FATAA through periodic consultations.
- The Office of Management and Budget (OMB), the agency tasked with establishing guidelines for implementing FAATA, should set a high bar for best practices in evaluation policies among the foreign assistance agencies. According to the legislation, OMB has until January of 2018 to set forth "guidelines - according to best practices of monitoring and evaluation studies and analyses - for the establishment of measurable goals, performance metrics, and monitoring and evaluation plans that can be applied with reasonable consistency to covered United States foreign assistance."

While it is important to ensure realistic standards for all the foreign assistance agencies, OMB should nevertheless set its standards high enough to produce credible and useful evaluations that actually inform decision-making. This means ensuring the guidelines include not only guidance on when to conduct evaluations, but also guidance on how to ensure evaluations are high quality - including guidance on how to ensure sampling techniques are adequate, relevant data is collected, data analyses are transparent and appropriate, alternate explanations for results are considered by the evaluators, and evaluators are required to disclose conflicts of interest in their work.

In addition, OMB should include guidelines for the dissemination of evaluations that go beyond posting completed evaluations online within 90 days of their completion, as required by FATAA. These could include recommended practices such as hosting post-evaluation meetings to discuss recommendations; producing written action plans for addressing the recommendations; sharing evaluations findings through summaries of findings, newsletters, or topic-specific platforms; hosting evidence summits or other events to share findings with relevant stakeholders (including partner country governments and beneficiaries); and including incentives for staff to prioritize learning from evaluations by ensuring it is an explicit part of their portfolio with adequate time built in and performance metrics that reward learning.

While the guidance should set a high bar and include specific guidance on ensuring evaluation quality and utilization, it should represent a "floor" and not a "ceiling". This means that as long as agencies meet the requirements in the guidance, they should also be encouraged to exceed and innovate beyond the parameters of the OMB guidance.

In addition, OMB should also ensure that agencies follow their own evaluation policy guidance and set aside resources for evidence-building activities (such as evaluations) both at the project level and at headquarters in their budget requests.

## Quality of Evaluations

- Agencies should increase the quality of evaluations by demanding higher standards for data sampling, collection, and analysis and by ensuring resources are available to meet these standards. In order to do this, agencies need to ensure their staff are knowledgeable about high standards of methodological rigor so that they can assess whether data collection tools and techniques, as well as data analysis methods, are sufficiently rigorous to produce relevant and credible findings.

It is also important to ensure resources are available to properly collect and analyze primary data. A lack of adequate resources is likely one of the reasons more than 40% of evaluations still do a poor job sampling and collecting data. When possible and relevant, evaluators should be encouraged to share and make public their primary data.

However, high quality evaluations do not have to be prohibitively expensive. The GAO report notes that in their sample there were "high quality" evaluations done for as little as $97,100. One interview respondent noted, "You don't necessarily have to spend more money to make your methodology more rigorous. If you only have money to survey people in three sites, that's fine, but randomize those sites instead of picking the ones that are most convenient."

- Agencies should consider prioritizing additional impact and ex-post evaluations, when appropriate. While impact and ex-post evaluations are not inherently more rigorous than performance evaluations, they do add distinct value to an overall evaluation portfolio. Impact evaluations allow donors to test unproven 'theories of change' before devoting large amounts of resources in interventions, and they help donors scale up proven interventions. Ex-post evaluations allow donors to assess the sustainability and long-term results of their projects. By prioritizing additional impact and ex-post evaluations when appropriate, agencies can ensure they have a more diverse and robust pool of information to inform their policies and programs.

## Accessibility and Disseminations of Evaluations

- Agencies and evaluators should work to make evaluations more accessible by increasing the publication of Summaries of Findings (as required by FATAA) and similar concise communications (such as one page Fact Sheets for senior policy makers) focused on learning and policy actions; ensuring wider and better distribution of evaluations and findings; and working to cluster and synthesize evaluations findings by sectors. While progress has been made in ensuring that evaluations are posted online, this is only the first step in encouraging transparency and utilization of evaluation findings.
- Agencies and evaluators should work with partner governments, local populations, and NGOs to ensure they have access to relevant evaluation findings by including them in evaluation design and implementation, sharing final evaluations, and working to determine how evaluation findings can inform future policies or projects. Currently, some agencies and implementers do a poor job of including partner governments and other stakeholders from developing countries in the evaluation process. Only about a third of USAID's evaluations include a local evaluator, and USAID's evaluations are shared with country partners less than a third of the time.

However, there are some cases of agencies coordinating with local stakeholders and partner country governments to ensure better use of evaluations. MCC includes local stakeholders in the evaluation design process, and then also includes both local stakeholders and partner governments in reviewing the final evaluations prior to publication. In one example MCC decided to invest additional funds in a randomized evaluation of a livestock management project when the Government of Namibia expressed explicit interest in using the findings from the evaluation to inform their own budgeting and programing in this area. "We decided to invest more than we might normally spend on evaluating a project of this size, because the findings very well might inform the government's decision-making and budget allocation," explained one respondent.

Agencies should make a point to do more of these types of research funds and evaluation partnerships, as well as expand efforts to build the capacity of local evaluators and proactively share evaluation findings with local stakeholders.

## Utilization of Evaluations for Accountability and Learning

- Congress should create more flexible budgets for agencies so that they may utilize evaluation findings and evidence to inform resource allocation. If agencies can demonstrate that they are, in fact, shifting funds at the project level based on what they have learned through evaluations, Congress should respond, allowing for more flexible budgets so that learning can take place at a larger, program level scale too.
- Similarly, agency leadership should request evaluation findings and evidence when approving project designs, either from their staff or from implementing partners. This should be done systematically, not on occasion. Prior to approving any project, leadership should ask the program designers how previous evaluations have been used to inform project design. This should become a regular part of every project approval discussion in order to make it clear that utilizing evaluation findings is a vital part of project design. If there are no prior evaluations, leadership should ensure the project has a clear development hypothesis and encourage the design of an evaluation strategy to ensure findings and lessons learned exist for future, similar projects.

Recognizing the importance of learning from previous evaluations, MCC's revised M&E policy now requires proposed project documentation to include a section on "how (evaluation) lessons have been reflected in the design and implementation of new interventions" and further requires leadership to verify the completion and relevance of these lessons to the proposed projects.

# When does it make sense to shift program funding based on evaluation findings?

One barrier to conducting and sharing independent, objective, and rigorous **performance evaluations** is fear that if problems are revealed or targets are not met, future funds might be cut off. In some cases - such as fraud, corruption, or rampant mismanagement of the project – cutting off funds may be the appropriate response. But in most cases, performance evaluations reveal issues in project design or implementation that can be improved upon in the next iteration. In order to incentivize rigorous and objective evaluations, it is important to preserve the space to reveal issues, learn, and adapt.

In contrast, **impact evaluation** findings are usually related to whether an intervention itself is successful. If there is sufficient evidence from impact evaluations that an intervention is not effective in meeting its objectives – even when it is implemented well – donors should stop funding that intervention and reallocate funds to interventions that are more successful in meeting their objectives.

- Program and operations staff should be more involved in designing scopes of work (SOWs) for evaluations, as well as the evaluation itself. In addition, there should be staff in agencies, implementing partners, and independent evaluators whose primary – if not sole – responsibility is ensuring evaluation findings are shared, internalized, and acted upon. Some agencies and organizations have already taken a lead on the latter recommendation. For example, over the past several years CARE has developed a position called Technical Advisor for Knowledge Management for various sectors. The position promotes "a culture of learning, information sharing, dialogue, and critical analysis" in order to "improve the quality of programming, adoption of good practices, and influencing policies and practices at scale." Staff in this position are encouraged to draw from evaluation and research both inside CARE and also from external partners and stakeholders. MCC also recently hired its first Director of Results and Learning to fill a similar role. By creating full-time positions dedicated to proactively learning and sharing information from evaluations, leadership can both solve a primary constraint to utilizing evaluation findings – the lack of staff time – and also signal the importance their organization puts on evidence-based programming and policy-making.
- Agency leadership should put systems in place to ensure evaluation recommendations are systematically shared with relevant stakeholders and a plan is put in place to respond to recommendations. Evaluation catalogs should be more user-friendly and include easy to navigate filters, including type of document (performance evaluation, impact evaluation, summary of findings, datasets, or others), project sector, and project country. Beyond online catalogues, agencies should also develop evaluation distribution plans. These should include proactively sharing the evaluations with program and operations staff, implementing partners, partner country governments, local aid recipients, and other stakeholders in both the US and the partner country. They should put in place systems that allow these stakeholders to respond to the evaluation findings and recommendations. In addition, specific steps should be in place for responding to evaluation recommendations, including standardized post-evaluation meetings of all relevant stakeholders and decision makers, as well as written action plans detailing if and how the agency or implementing partner plans to respond to recommendations.

Agency leadership should develop a set of concrete examples for use internally and with Congress and outside stakeholders, in order to demonstrate the critical role that evaluation and learning play in supporting effective foreign assistance. These examples should demonstrate how learning enabled their agencies to make course corrections, close unsuccessful programs, stand up more effectives ones, and change policy. This should be done as more than just a public relations exercise – it should be a tool that agencies use to make difficult decisions about where to best allocate scarce funds.

- OMB should serve as a consumer of evaluation findings by asking agencies how evidence and evaluations inform their budget requests. By linking both evaluation resources and evaluation findings into the budget build process, OMB can help ensure evaluations are utilized to achieve the most effective outcomes.

# 06 Conclusion

When Congress passed the Foreign Aid Transparency and Accountability Act in 2016, it did so with broad bipartisan support, passing without objection in both the House and Senate. Civil society advocates, including MFAN, had strongly supported the legislation, as they saw it as a way to lock in foreign assistance reforms in transparency and evaluation. Even staff at the foreign assistance agencies – many of whom often view congressional requirements as burdensome – welcomed FATAA as an important piece of legislation that safeguards their recent progress.

This widespread, bipartisan support across policymakers, practitioners, and advocates reflects a shared agreement that effective foreign assistance programs strengthen America's standing in the world, bolster its national security, and save or improve the lives of millions of people around the world. However, a critical component for ensuring this assistance is as effective as possible is through producing and utilizing high-quality evaluations.

"There is ample evidence that aid can be tremendously effective and similar evidence that evaluations can play a vital role in ensuring this effectiveness."

**There is ample evidence that aid can be tremendously effective and similar evidence that evaluations can play a vital role in ensuring this effectiveness.** In "Millions Saved," public health experts demonstrated this by compiling the findings of dozens of rigorous evaluations in public health to highlight what worked – and what didn't – in global health projects. As the title suggests, they found some projects made significant, measurable and positive impacts on peoples' lives. These ranged from a project that developed a low-cost vaccine for meningitis A, which was successfully rolled out to 217 million people in four years (virtually eliminating meningitis in the areas the project covered) to reducing traffic fatalities by incentivizing motorcyclist to wear helmets.

Similarly, the Poverty Action Lab's policy lessons use evaluation findings to highlight how, why, and when development projects work in various sectors. Among the many projects they evaluate, one focused on transferring an asset (such as livestock) to the poorest families and providing supplemental training and support for maintaining the asset. Looking across seven versions of this

project in seven countries, they found the asset transfer and support resulted in increased consumption, food security, asset holding and savings among these ultra-poor families, even three years after the initial asset transfer. The program's benefits exceeded its costs in six out of seven countries, resulting in a positive rate of return ranging from 133 to 433 percent. These types of evaluations and analyses demonstrate cost-effective methods for saving and improving the lives of millions.

Evaluation policies and practices within USG foreign assistance agencies have come a long way in recent years. However, this report demonstrates that there is still more work that needs to be done to transition agencies from fully implementing their own evaluation policies and adhering to FATAA provisions to developing an ingrained culture that values evaluations as a means for learning and implementing what works. Such a culture would ensure the quality of evaluation methodologies. The new administration, Congress, the advocacy community, and leadership and staff within these agencies all have a role to play in ensuring the USG is a leader in providing effective, evidence-based foreign assistance, and the utilization of evaluation is a critical tool to achieving this goal. The most effective foreign aid programs can serve as key building blocks toward a more safe, stable and prosperous world.

# Annex
Interview Questions

01

**Evaluation Policy**

- What were the motivations for adopting an evaluation policy at your agency (or the agency you work with most often?)
- What types of discussions occurred around drafting the policy? Were there varying views on what it should contain?
- Once the evaluation policy was adopted, what types of changes occurred in your agency (of the agency you work with most often?)
- What types of changes were most difficult to implement? What were the hurdles? How were they overcome?
- Five years later, what progress has your agency made in the evaluation space? Where have you not seen the progress you may have hoped to achieve?

**Quality of Evaluations**

- How would you personally define a high quality evaluation? What aspects of quality are most important to you?
- What are the main obstacles to getting high quality evaluations?
- How does your organization determine when to employ what type of evaluation?
- How "ex-post" can an evaluation be and still capture causation (or even strong correlation?)
- How do you think the quality of your evaluations compares to five or ten years ago?

**Use of Evaluations**

- Who are the intended users of your evaluations? How do you hope they will be used?
- Do you believe they are currently being used for those purposes?
- If so, how has that resulted in more effective development practices?
- If not, what are the main barriers for use?
- How could those barriers be overcome?

**Lessons Learned and Recommendations**

- In the past five to ten years, how has your agency improved its evaluation practices? What were the drivers of change?
- What obstacles have proven very challenging to overcome? What advice would you give to others trying to overcome the same set of obstacles?

NB: Not all interviewees were asked all questions. Questions were tailored to the individual based on their experiences, expertise, and time availability.

# Annex
Survey Questions

02

**US Government Staff (Current or Former)**

**Section 1: Personal Information**

Name:

Agency or Organization:

Your primary job is (pick best fit):

- M&E Specialist (USG agency)
- Program Design/Implementation (USG agency)
- Both M&E and Program Design/Implementation (USG agency)
- Communications, Public Affairs, Congressional Affairs (USG agency)
- M&E Implementer (Non-profit or private implementer)

  Please write-in the agency you work with the most often on evaluations
- Program Design/Implementation (Non-profit or private implementer)

  Please write-in the agency you work with the most often on design or implementation

**Section 2: Evaluation policy**

Did your agency (or the agency you work for most often) have any of the following concerns about implementing its evaluation policy (Check all the apply)

- Evaluations might reveal shortfalls in programs or operations
- Evaluations might be too expensive
- Evaluations might be too time consuming
- Evaluation findings might not be used
- Other concerns (please write in)
- No concerns

At your agency (or the agency you work for most often), has there been more, less, or the same amount of the following since its evaluation policy was adopted:

- Clarity as to the roles and responsibilities of staff regarding evaluation?
- Leadership support for devoting time and energy to conducting evaluations?
- Staff who are appropriately trained and skilled at designing, managing, and assessing evaluations?
- Budget allocated to evaluations?
- High quality SOW for evaluations?
- Completion of high quality performance evaluations?
- Completion of high quality of impact evaluations?
- Leadership pressure to use evaluations to inform ongoing or future projects?
- Actual use of performance evaluations to inform ongoing or future projects?
- Actual use of impact evaluations to inform ongoing or future projects?

On a scale of 1 – 5, with 1 being no impact and 5 being huge impact, I would rate the impact of the evaluation policy in improving the quantity, quality, and use of evaluations at my agency (or the agency I work for the most often) as:

- Improving quantity
- Improving quality
- Improving the use of evaluation findings or recommendations

Please describe the top 2-3 key issues your agency (or the agency you work with most often) met while trying to implement the evaluation policy. How, if at all, were these issues resolved? (Open ended)

**Section 3: State of Evaluations and evaluation quality**

When it comes to evaluation findings, I believe the following incentives are in place to encourage transparency (scale of 1 to 5 for each, with 1 being "strongly disagree that these incentives are in place" and 5 being "strongly agree.")

- My organization believes it will be rewarded by stakeholders (including civil society and Congress) if it is fully transparent about evaluation findings – even negative findings.
- My organization believes that conducting high quality evaluations and publishing their findings as transparently as possible is a public good and a priority.
- My organization's leadership encourages staff to flag both negative and positive findings from evaluations so that the organization can improve its program and project design.
- Staff at my organization feel safe designing innovative or experimental projects, along with evaluations to test the effectiveness of the innovation.
- Staff at my organization feel safe revealing negative findings from an evaluation to their peers and managers.

Currently, staff working ay my agency (or the agency I work for most often) have adequate resources in the following areas (Rate on scale of 1 – 5, with 1 being "almost no resources" and 5 being "totally sufficient resources")

- Budget for evaluations
- Staff time and expertise to manage evaluations
- Access to high quality independent evaluators
- Leadership who encourage transparent and proactive communication of evaluation findings
- Staff time and expertise to communicate evaluation findings to internal or external stakeholders
- Leadership who encourage and reward evidence-based program design
- Staff time and expertise to incorporate evaluation findings into programing

Does your agency (or the agency you work for most often) have any of the following concerns about implementing impact evaluations, specifically? (Check all the apply)

- Impact evaluations might reveal shortfalls in programs or operations
- Impact evaluations are too expensive
- Impact evaluations are too time consuming
- Impact evaluations are not often designed or implemented in a way that gives us the information we need
- Impact evaluation findings are not used often by internal or external stakeholders
- Most our projects are not amenable to experimental design
- Other concerns (please write in)
- No concerns

(USG only) In my opinion, evaluations in my team or bureau provide (Often, sometimes, rarely, or never):

- Information on our projects' implementation, results, or sustainability that I did not previously know
- Evidence of impact on beneficiaries (or lack thereof) that can be directly attributed to our projects
- Specific findings or recommendations that help me to design better projects in the future

**Section 4: Evaluation Usage**

In my agency or organization, I believe the following incentives are in place to encourage evaluation use (scale of 1 to 5 for each, with 1 being "strongly disagree that these incentives are in place" and 5 being "strongly agree.")

- The use of evaluation findings is required or strongly encouraged in strategic reports and program design.
- Programs or project are generally not approved unless they explicitly incorporate lessons learned from previous similar projects and/or there is sufficient empirical evidence that they are likely to be effective.
- Leadership actively encourages transparency and accountability through their actions (including encouraging course corrections and rewarding staff who point out project shortcomings revealed by evaluations.)
- Leadership actively encourages staff to take time to reflect, learn, and incorporate learning into future programs (including through work prioritization and staff performance reviews.)
- There is a culture or learning and/or accountability because staff and management believe this is a fundamental part of our organization's mission.

I have used findings or recommendations from the following types of evaluations to inform ongoing or future project design or implantation:

- A performance evaluation I was involved with designing, managing, or assessing
- An impact evaluation I was involved with designing, managing, or assessing
- A performance evaluation I was not involved with in any way
- An impact evaluation I was not involved with in any way

Are you aware of anyone outside of your agency or organization doing any of the following:

- Reading an evaluation completed by your organization or published by your agency
- Using findings from your evaluation to disseminate knowledge
- Using findings from your evaluation to inform the design of a subsequent project

Please describe any examples of evaluation findings being used by yourself, colleagues, or external stakeholders (open ended)

Do you think any of the following serve as obstacles to using findings from the agencies' evaluations (rate on a scale of 1 to 5 with 1 being "not an obstacle" and 5 being "very big obstacle"):

- The evaluations are not public, easy to find, or otherwise accessible to most stakeholders
- The quality of the evaluation is not sufficient to use its findings
- The evaluations are too long or complex to use
- The evaluation findings are not generalizable to other projects
- There is not sufficient interest or time during project design or implementation to read and internalize evaluation findings
- Other (please describe)

What is needed to increase the internal use of evaluation findings at your organization or agency? (Open ended)

What is needed to increase the external use of evaluation findings at your organization or agency? (Open ended)

**Section 5: The role of civil society**

To what extent did the efforts of civil society, including researchers, advocacy groups or think tanks, influence your agency's efforts to improve the quantity, quality or use of evaluations? (Scale of 1 to 5 with 1 being "not at all" and 5 being "a major influence")

- Quantity of performance evaluations
- Quantity of impact evaluations
- Quality of performance evaluations
- Quality of impact evaluations
- Use of performance evaluations
- Use of impact evaluations

To what extent did MFAN specifically provide an incentive to improve transparency efforts in your organization? (Scale of 1 to 5 with 1 being "not at all" and 5 being "major incentive.")

To what extent did the following civil society efforts impact your organization's evaluation efforts (Scale of 1 to 5 with 1 being "very negative impact", 3 being "no impact", and 5 being "very positive impact")

- Information or opinion pieces about the benefits of evaluations
- Information or opinion pieces that were critical of your agency/US efforts on evaluations
- Information or opinion pieces the praised your agency's efforts on evaluations
- Technical assistance (i.e. executive trainings, technical discussions, etc.)

In your opinion, what civil society efforts were most helpful to furthering evaluation policy in the US? Which efforts were least helpful or even counter-productive? (Open ended.)

**Section 6: Lessons Learned and Recommendations**

In your opinion, how successful have efforts to improve the quality and use of evaluation been in your agency (or the agency you work for most often) in the past five years (Scale of 1-5 with 1 being totally unsuccessful and 5 being very successful) ___

- Improving the quality of performance evaluations
- Improving the quality of impact evaluations
- Improving the use of performance evaluations to inform ongoing or future projects
- Improving the use of impact evaluations to inform ongoing or future projects

To the extent that efforts have been successful, which of the following were most important for your agency's success (Score each on a 1-5 scale with 1 being "not very important" and 5 being "crucial")

- Adoption of an Evaluation Policy
- Increased leadership support and interest in evaluation findings

- Ensuring sufficient budget and financial resources
- Ensuring sufficient staff time and expertise
- Hiring evaluators who have sufficient expertise in evaluations
- Other (please write in)

To the extent that efforts have been challenging, what factors serve as the biggest constraints (Score each on a 1-5 scale with 1 being "not a constraint "and 5 being "a major constraint")
- Fear of evaluations showing poor results
- Insufficient leadership support and/or interest in using evaluation findings
- Insufficient budget / financial resources
- Insufficient staff time and expertise
- Hiring evaluations with insufficient expertise in evaluation
- Other (please write in)

Based on your experiences with evaluations, what are your top 2-3 lessons learned? What recommendations would you make to improve the process and the outcomes in the future? (Open ended)

Is there anything else you want to share about your experiences or recommendations regarding evaluations and their use? (Open ended)

**Civil society**

When advocating for improving evaluations of US foreign assistance, what arguments about the benefits of evaluation did you find were compelling to agencies or organizations? (Score each on a scale of 1 – 5 where 1 = not compelling and 5 = very compelling)
- Evaluations can help ensure projects are being implemented as intended and targets are being reached
- Evaluations can help test theories and build evidence on what works in international development
- Evaluations can help donors communicate the impact of their work to stakeholders
- Evaluations can help donors ensure money is being well-spent
- Evaluations can help donors and implementers design better projects and programs in the future
- Other (please write in)
- None of the above

When advocating for evaluations, what are the main sources of resistance from agencies or organizations? (Check all that apply)
- Evaluations might reveal shortfalls in programs or operations
- Evaluations might be too expensive
- Evaluations might be too time consuming
- Evaluation findings might not be used
- Evaluation results might be used as an argument against US foreign assistance programs
- Other concerns (please write in)
- No concerns

Do you believe most US development agencies have sufficient resources in the following areas: (Please answer yes, no, somewhat, or not observable)

- Budget for evaluations
- Staff time and expertise to manage evaluations
- Access to high quality independent evaluators
- Leadership who encourage transparent and proactive communication of evaluation findings
- Staff time and expertise to communicate evaluation findings to internal or external stakeholders
- Leadership who encourage and reward evidence-based program design
- Staff time and expertise to incorporate evaluation findings into programing
- How successful would you rate the US government's efforts to increase the quantity and quality of evaluations over the past five years? (Scale of 1-5 with 1 being totally unsuccessful and 5 being very successful)
- How successful would you rate the US government's efforts to increase the use of evaluations over the past five years? (Scale of 1-5 with 1 being totally unsuccessful and 5 being very successful)
- Are there agencies you believe have done a particularly good job? What about a particularly poor job? (Open ended.)
- How successful would you rate your organization's advocacy efforts to increase the quality and quantity of evaluations over the past five years? (Scale of 1-5 with 1 being totally unsuccessful and 5 being very successful)
- Based on your experiences with promoting more and better evaluations (and their use), what recommendations would you make to improve the quality and use of evaluations in the future? (Open ended)
- Is there anything else you want to share about your experiences or recommendations regarding improving evaluations of US foreign assistance? (Open ended)

**Evaluators**

At the agency you work with the most often, has there been more, less, or the same amount of the following since its evaluation policy was adopted: (add an unobservable answer)

- Demand for high quality performance evaluations?
- Demand for high quality of impact evaluations?
- Agency staff who are appropriately trained and skilled at designing, managing, and assessing evaluations?
- Budget allocated to evaluations?
- High quality SOW for evaluations?
- Use of performance evaluations to inform ongoing or future projects?
- Use of impact evaluations to inform ongoing or future projects?

When it comes to evaluation findings, I believe the following are true for the agency I work with most often (scale of 1 to 5 for each, with 1 being "strongly disagree that these incentives are in place" and 5 being "strongly agree.")

- The agency is open and receptive to critical evaluation findings.
- Staff and management appear eager to learn from evaluation findings.
- The agency believes in publishing evaluation findings in a transparent manner.

Are you aware of anyone within the agency you work with most often doing any of the following:

- Proactively sharing your evaluations with country partners or beneficiaries
- Using findings from your evaluation to disseminate knowledge about what works and what doesn't in the sector of interest
- Using findings from your evaluation to inform the design of a subsequent project

Please describe any examples of evaluation findings being used by the agency or external stakeholders (open ended)

In your opinion, how successful have efforts to improve the quality and use of evaluation been in the agency you work for most often in the past five years (Scale of 1-5 with 1 being totally unsuccessful and 5 being very successful)
- Improving the quality of performance evaluations
- Improving the quality of impact evaluations
- Improving the use of performance evaluations to inform ongoing or future projects
- Improving the use of impact evaluations to inform ongoing or future projects

To the extent that efforts have been successful, which of the following were most important for the agency's success (Score each on a 1-5 scale with 1 being "not very important" and 5 being "crucial")
- Adoption of an Evaluation Policy
- Increased leadership support and interest in evaluation findings
- Ensuring sufficient budget and financial resources
- Ensuring sufficient staff time and expertise
- Strengthening in-house staff capacity on evaluation
- Hiring evaluators who have sufficient expertise in evaluations
- Other (please write in)

To the extent that efforts have been challenging, what factors serve as the biggest constraints (Score each on a 1-5 scale with 1 being "not a constraint "and 5 being "a major constraint")
- Fear of evaluations showing poor results
- Insufficient leadership support and/or interest in using evaluation findings
- Insufficient budget / financial resources
- Insufficient staff time and expertise
- Hiring evaluators with insufficient expertise in evaluation
- Other (please write in)

Based on your experiences working with USG on evaluations of foreign assistance, what do you think could improve the quality and use of evaluations of foreign assistance in the future? (Open ended)

Is there anything else you want to share about your experiences or recommendations regarding evaluations and their use? (Open ended)

**Implementers**

Over the past 5 years, have you seen evidence that the agency you work with most often is making mid-course corrections to project implementation based on evaluation findings? (Often, occasionally, rarely, never, don't know.)

Over the past 5 years, have you seen evidence that the agency you work with most often is incorporating evaluation findings into subsequent designs for program or projects? (Often, occasionally, rarely, never, don't know.)

If so, how are evaluation findings changing program design or implementation? (Open ended)

Can you give specific examples of evaluation findings being used to make mid-course corrections or influencing subsequent project design?

Does the agency you work with most often actively encourage evidence-based project design? (Often, occasionally, rarely, never, don't know.)

Has your organization used evaluation findings to inform program and project design? (Yes, occasionally, no, don't know.)

If so, can you share examples?

Based on your experiences working with USG on evaluations of foreign assistance, what do you think could improve the quality and use of evaluations of foreign assistance in the future? (Open ended)

Is there anything else you want to share about your experiences or recommendations regarding evaluations and their use? (Open ended)

# Annex

## Findings on Evaluation Quality from "Meta-Evaluation of Quality and Coverage of USAID Evaluations"

03

- Questions were linked to purpose (100% -> 98%)
- Data collection methods described (92% -> 96%)
- Project characteristics described (90% -> 91%)
- Social science methods (explicitly) were used (81% -> 84%)
- Annex included list of sources (84% -> 83%)
- External team leader (64% -> 82%)
- Management purpose described (70% -> 81%)
- Annex included data collection instruments (56% -> 81%)
- Findings supported by data from range of methods (68% -> 80%)
- Recommendations clearly supported by findings (80% -> 79%)
- Recommendations specific about what is to be done (58% -> 77%)
- Project "Theory of Change" described (77% -> 74%)
- Evaluation questions addressed in report (not annexes) (59% -> 74%)
- SOW is included as a report annex (45% -> 74%)
- Questions in report same as in SOW (12% -> 69%)
- Findings are precise (not simply "some, many, most") (74% -> 67%)
- Study limitations were included (38% -> 64%)
- Recommendations—not full of findings, repetition (58% -> 64%)
- Report structured to respond to questions (not issues) (47% -> 51%)

---

- Findings distinct from conclusions/recommendations (37% -> 48%)

---

- Executive summary mirrors critical report elements (41% -> 45%)
- Recommendations specify who should take action (43% -> 45%)
- Report discusses differential access/benefit for men/women (42% -> 40%)
- Evaluation team included local members (33% -> 35%)
- Data analysis method described (34% -> 34%)
- Number of evaluation questions was 10 or fewer (49% -> 29%)
- Data collection methods linked to questions (17% - 22%)
- Evaluation findings sex disaggregated at all levels (23% -> 22%)
- Data analysis methods linked to questions (32% -> 19%)
- Report said team included an evaluation specialist (15% -> 19%)
- Unplanned/unanticipated results were addressed (15% -> 14%)
- Written approval for changes in questions obtained (8% -> 12%)
- Report indicated conflict-of-interest forms were signed (0% -> 12%)
- Alternative possible causes were addressed (10% -> 10%)
- Reason provided if some questions were not addressed (21% -> 9%)
- Evaluation SOW includes Evaluation Policy Appendix 1 (0% -> 8%)
- Statements of differences included as an annex (3% -> 7%)
- Report explains how data will transfer to USAID (0% -> 5%)

# Footnotes

1 Strengthening Evidence-Based Development. Five Years of Better Evaluation Practice at USAID (2011 – 2016). USAID. March 2016.

2 Review of Department of State Compliance with Program Evaluation Requirements. Office of the Inspector General. September 2015.

3 The GAO report does not assess the quality of PEPFAR evaluation as a comprehensive initiative. However, it does assess PEPFAR evaluations conducted by the Department of Health and Human Service's Center for Disease Control and Prevention (CDC.)

4 Meta-Evaluation of Quality and Coverage of USAID Evaluations: 2009-2012. Management Systems International. August 2013. See page 156 for more details on how MSI constructed its composite quality score.

5 The lack of clarity often comes from stakeholders who discuss "evaluations" without making the distinction between different purposes, types, or methods of evaluations. It is worth noting that all the agencies covered in this assessment have slightly different language in their policies on evaluation purposes, types, and definitions, but they all clearly define performance and impact evaluations as two distinct types of evaluation and encourage staff to select the evaluation methodology that is most appropriate for answering the relevant questions.

6 Evaluation Utilization at USAID. Management Systems International. February 23, 2016.

7 Beyond Success Stories: Monitoring and Evaluation for Foreign Assistance Results, Evaluator Views of Current Practice and Recommendations for Change. Richard Blue, Cynthia Clapp-Wincek and Holly Benner. May 2009.

8 Evaluation: Learning from Experience. USAID Evaluation Policy. January 2011.

9 The Biggest Experiment in Evaluation: MCC and Systematic Learning. William Savedoff, Center for Global Development. November 2012.

10 Leading Through Civilian Power. U.S. Department of State - Quadrennial Diplomacy and Development Review. 2010.

11 Evaluation: Learning from Experience. USAID Evaluation Policy. January 2011.

12 Large projects are defined as those whose costs equal or exceed the mean value of projects for the OU.

13 Strengthening Evidence-Based Development. Five Years of Better Evaluation Practice at USAID (2011 – 2016). USAID. March 2016.

14 Department staff refute this finding. They note "The OIG figure included bureaus that do not do foreign assistance. The number of foreign assistance bureaus that did not meet the requirement was closer to 27% in 2014, and some of those 27% had completed one but not two evaluations."

15 Review of Department of State Compliance with Program Evaluation Requirements. Office of the Inspector General. September 2015.

16 Impact Evaluations in Agriculture: An Assessment of the Evidence. Independent Evaluation Group, World Bank.. 2011.,

17 MCC's First Impact Evaluations: Farmer Training Activities in Five Countries. MCC Issue Brief. October 2012.

18 The 2012 Evaluation Policy technically only lasted for two years; it was not a permanent policy. So the 2015 revision was not just an "update" – it prevented the evaluation policy from disappearing altogether.

19 Meta-Evaluation of Quality and Coverage of USAID Evaluations: 2009-2012. Management Systems International. August 2013. See page 156 for more details on how MSI constructed its composite quality score.

20 The GAO report does not assess the quality of PEPFAR evaluation as a comprehensive initiative. However, it does assess PEPFAR evaluations conducted by the Department of Health and Human Service's Center for Disease Control and Prevention (CDC.)

21 The GAO report actually assesses 173 evaluations across six agencies. Only data from USAID, State, MCC, and HHS are pulled for this assessment; data from the Department of Defense Global Train and Equip program and the US Department of Agriculture Foreign Agricultural Service are not included.

22 Data was considered "most relevant" if it was collected from at least two of the following subjects (if relevant): Program recipients, program non-recipients, and goods and services AND it tracked at least two of the following: Whether people used the goods and services provided, whether they demonstrated knowledge or behavior changes based on training, and whether there were changes in health outcomes. If secondary data was used, it was assessed on whether the data was relevant to the program being evaluated.

23 How Well Are Aid Agencies Evaluating Programs? An Assessment of the Quality of Global Health Evaluations. Julia Goldberg Raifman, Felix Lam, Janeen Madan, Alexander Radunsky, and William Savedoff. Forthcoming.

24 A number of interviewees noted that the presence of a conflict of interest form may not be the best proxy for assessing whether a conflict of interest is likely. There are other steps agencies take to prevent conflict of interest. For example, at MCC the standard evaluator contract language requires any potential conflicts of interest to be fully documented in the published evaluation report. Similarly, in USAID's updated ADS 201, disclosures of conflict of interest are a required component in USAID's evaluation reports.

25 Evaluation Utilization at USAID. Management Systems International. February 23, 2016.

26 Although there have not been similar external publications, current and former staff at MCC indicate efforts to aggregate and synthesize evaluation findings are pending based on impact evaluations of roads and performance evaluations of anti-corruption projects.

27 The 2017 US foreign aid budget and US global leadership: The proverbial frog in a slowly heating pot. George Ingram, Brookings. Feb 2016.

# Photo Credits

Cover –
Courtesy of Paula Bronstein/Getty Images Reportage. Some rights reserved.

Pg 9 – Right
Courtesy of Juan Arredondo/Getty Images Reportage. Some rights reserved.

Pg 12 – Left
Courtesy of Jonathan Torgovnik/Getty Images Reportage. Some rights reserved.

Pg 15 – Right
Courtesy of Juan Arredondo/Getty Images Reportage. Some rights reserved.

Pg 21 – Right
Pixabay.com. Some rights reserved.

Pg 25 – Right
Pixabay.com. Some rights reserved.

Pg 28 – Left
Courtesy of Juan Arredondo/Getty Images Reportage. Some rights reserved.

Pg 34 – Left
Courtesy of Jonathan Torgovnik/Getty Images Reportage. Some rights reserved.

Pg 41 – Right
Courtesy of Paula Bronstein/Getty Images Reportage. Some rights reserved.

Pg 46 – Left
Courtesy of Jonathan Torgovnik/Getty Images Reportage. Some rights reserved.

Pg 51 – Right
Pixabay.com. Some rights reserved.

Pg 59 – Right
Courtesy of Juan Arredondo/Getty Images Reportage. Some rights reserved.