

# Six Key Issues to Consider when Developing an Evaluation Statement of Work

*Discussion Paper*

*Jerome Gallagher, USAID/PPL/LER*

*7/1/13*

In the course of reviewing many evaluation Statements of Work (SOWs), there are a few issues that come up again and again. Here are six key issues to keep in mind when developing an evaluation SoW. This is not an exhaustive list, but hopefully a useful one.

## **1. Sector assessments are not evaluations; expert reviews are not evaluations.**

Evaluation is not the only type of research that USAID conducts; nor is it the only type of value. Two other types of studies that USAID conducts often get confused with evaluations - assessments (or “sector assessments”) and informal reviews. Here’s the definition of *Evaluation*:

*Evaluation* is the systematic collection and analysis of information about the characteristics and outcomes of programs and projects as a basis for judgments, to improve effectiveness, and/or inform decisions about current and future programming.

As the evaluation policy goes on to note:

Evaluation is distinct from *assessment*, which may be designed to examine country or sector context to inform project design, or an *informal review* of projects.

USAID evaluations focus on USAID interventions; sector assessments focus on the entire sector in which our interventions operate. Sector assessments tend to be more forward looking with an emphasis on what the needs are that should be addressed with our programming, while evaluations tend to be more backward looking regarding our efforts in addressing particular needs. Both sector assessments and evaluations are important and both can inform future programming; they are just different. Evaluations must meet the standards of the evaluation policy, while this is not true of sector assessments.

There is no reason why a SoW can’t combine a sector assessment and an evaluation. One area where there are particular synergies in combining a sector assessment with an evaluation is in the development of recommendations. Combining information from a sector assessment and an evaluation can improve the recommendations made by an evaluator. However, combining a sector assessment and an evaluation should not be viewed primarily as a means of saving costs.

Informal reviews (I sometimes refer to them more charitably as “expert reviews”) are also not evaluations. There is no definition of “informal review” at USAID, but it generally refers to an

informal, subjective review of a project, typically by an expert in the field. Like sector assessments, informal reviews can be quite helpful to missions, particularly when they provide a time-sensitive check on project design and implementation from a trusted source, such as a USAID technical expert.

While informal reviews tend to be subjective, evaluations aim for objectivity. As the Evaluation policy notes, USAID evaluations are expected to involve: “Use of data collection and analytical methods that ensure, to the maximum extent possible, *that if a different, well-qualified evaluator were to undertake the same evaluation, he or she would arrive at the same or similar findings and conclusions,*” and “application and use to the maximum extent possible of social-science methods and tools that *reduce the need for evaluator-specific judgments.*”

If you are writing a SOW that looks like it will result in an evaluation that will be highly dependent on not just the skills of the evaluator, but also the evaluator’s preferences about programing in a particular sector, than it might be worth revisiting the evaluation questions and methodology section of the evaluation SoW.

## **2. Don’t call your evaluation an “Impact evaluation” just because you are interested in project impact.**

There are two main categories of evaluations at USAID - Impact evaluations and Performance evaluations. Here are the definitions:

*Impact evaluations* measure the change in a development outcome that is attributable to a defined intervention; impact evaluations are based on models of cause and effect and require a credible and rigorously defined counterfactual to control for factors other than the intervention that might account for the observed change. Impact evaluations in which comparisons are made between beneficiaries that are randomly assigned to either a treatment or a control group provide the strongest evidence of a relationship between the intervention under study and the outcome measured.

*Performance evaluations* focus on descriptive and normative questions: what a particular project or program has achieved (either at an intermediate point in execution or at the conclusion of an implementation period); how it is being implemented; how it is perceived and valued; whether expected results are occurring; and other questions that are pertinent to program design, management and operational decision making. Performance evaluations often incorporate before-after comparisons, but generally lack a rigorously defined counterfactual.

The Evaluation SoW should identify which type of evaluation you are requesting from the evaluation team. There has recently been a greater emphasis on Impact evaluations at USAID and it makes sense to want an evaluation that will tell you the extent to which your project has had an impact. It’s important to note, though, that what makes an evaluation an “impact evaluation” is not the intent of

the evaluation requester, it is the methodology used. If your SOW asks impact evaluation questions and (most importantly) it requests an impact evaluation methodology, in particular an experimental or quasi-experimental design that includes a rigorously defined counterfactual, then you can call it an impact evaluation in your SOW.

If you don't ask impact evaluation questions or if you don't request impact evaluation methodologies then you are probably requesting a performance evaluation. Please note that impact evaluations may include performance evaluation questions in addition to impact evaluation questions. Many impact evaluations, for instance, will also address questions regarding project implementation in addition to project impact. As long as it includes an impact question and methodology, then it is still an impact evaluation regardless of the other performance evaluation questions that are included in the SOW.

If you are planning to prepare an impact evaluation SoW, my best advice is to plan early (project design stage) and get help from a Washington expert. Because of the technical nature of impact evaluations, assistance from an impact evaluation specialist is critical for ensuring that the considerable resources that will be spent on an impact evaluation will lead to a high quality estimate of the impact of your project on defined outcomes.

Please note that a performance evaluation *can* also address project impact (just not in the rigorous manner of an impact evaluation). There is some confusion over the term "performance evaluation" that leads some to think that a performance evaluation should only focus on the "performance" of the implementer. That is not the case. The best way of thinking about "performance evaluations" is to think of them as any kind of evaluation that is not an impact evaluation. As the definitions above hopefully make clear, the definition of an impact evaluation is quite narrow, while the definition of a performance evaluation is far more broad. Performance evaluations can employ a wide range of methodologies to address questions ranging from descriptive, to normative, to cause-and-effect questions. For instance, a performance evaluation that employs case study methodology can certainly provide strong evidence that our project had an impact on a development outcome (but it is unlikely to quantify that impact and it won't employ a rigorous counterfactual as required for an impact evaluation).

### **3. If you do nothing else, make sure your evaluation questions are useful, limited, clear, and researchable.**

The evaluation questions are the core of an evaluation statement of work and affect every other section of the SOW. If there is one thing to get right about an evaluation SOW, it is the set of evaluation questions. Even more than the evaluation purpose and evaluation methodology, the questions will determine the content of the final report. As your external evaluators take what is in the SOW and move through the evaluation process (from designing and implementing the evaluation to writing the final evaluation report), the evaluation purpose will have less operational importance

and the methodology is likely to go through changes and adaptations, but the evaluators have an obligation to answer *every* evaluation question in the SOW, so it's critical that the SOW gets the questions right.

There are three critical qualities of evaluation questions that SOWs at USAID often flout, but before I get to those, let me just mention one overarching quality of good evaluation questions:

**Useful.** Ultimately, evaluations are to be used, so ask questions whose answers will help you manage your project or portfolio or for making a specific management decision. That requires reaching out to the primary audience of the evaluation. Since evaluation questions need to be limited, only include those questions whose answers will add the most marginal utility. The team that peer-reviews a SOW won't necessarily know what is useful to the primary users of the evaluation, so ensuring usefulness is really in the hands of the drafters of the SoW. But this doesn't mean that "usefulness" should be a reason for the drafters of the SOW to flout the other qualities mentioned below.

OK, now the three critical qualities of evaluation questions that SOWs at USAID often flout:

**i. Limited.** The USAID guidance on developing high quality SOWs recommends 3 to 5 evaluation question. Yet, most evaluation SOWs at USAID still include far more questions. It's not uncommon to see USAID performance evaluations with 10 to 20 or even more evaluation questions. Including more than five evaluation questions in you SOW is likely to cause big problems down the road leading to an evaluation report that covers too many issues without sufficient evidentiary depth to support credible answers to the evaluation questions. Of course, the bigger the budget you have and greater the LOE of the evaluators, the more questions you can potentially answer, but given the typical budget and LOE of USAID performance evaluations, three to five questions or is more than enough. Three to five questions also fits well with the expected length of evaluation reports. Recent guidance suggests that the "Findings, Conclusions, and Recommendations" section of an evaluation report should be 15 to 25 pages. (See the How-to Note on Preparing Evaluation Reports). If a SOW includes 25 questions, then the evaluator will have, at most, one page to present evidence and answer each evaluation question. That's just not enough. (And don't think that giving the evaluator carte blanche to write long Annexes will save you. It won't.)

It's worth considering GAO performance audits as a useful counterpoint to USAID performance evaluations. GAO performance audits tend to focus on only three to four questions despite far greater LOE (particularly in the analysis and reporting stage of the evaluation) compared to USAID performance evaluations that address a far greater number of questions. Check out some of their reports at <http://www.gao.gov>. Really, go check them out. Do you want examples? OK, here's one example: <http://www.gao.gov/products/GAO-12-728> about a USAID funded road project in Indonesia. It's a forty-five page report that asked three questions, examined a single 91 mile road project, and took six months to complete. If this was a USAID evaluation, I can't image that the SOW would have had only three questions or that we would have provided nearly as much resources. Do we really

think our external evaluators are so much better than GAO?

Note, too, that being limited in your evaluation questions is not just being limited in the number of questions, but being limited in the scope of the questions. It's about being focused on what is most important, what information gives you the most value added, and what is most achievable by the evaluators. Many USAID evaluations attempt to address seemingly every aspect of the project under review. This is not just an issue of having too many questions. An evaluation could have a single question and still have a scope that is too wide, e.g. "How was this program implemented and what were the outcomes and impacts across the different program objectives?"

It's fine for an evaluation to address a particular aspect of a project - a puzzle or truly unanswered question about the project that required evaluative research. It should not be the expectation at USAID that most evaluations are supposed to cover all of a project's or implementing mechanism's objectives so that an overall judgment about a large and multi-dimensional project can be made. That can be a worthwhile approach to evaluation in some instances, but given the limited resources put into many performance evaluations, USAID performance evaluations should be more narrowly focused.

**ii. Clear.** The meaning of *every word* in the evaluation question must be clear, particularly for abstract concepts like "objective" "effective" "processes". If it is not clear in the question itself, then you should provide a statement along with the question explaining what you mean. General definitions are not good enough; you need to explain what the word means in the particular context of the evaluation and/or the project being evaluated.

As examples, let's discuss "objective" and "effective". Here's a fairly typical evaluation question one sees in a SOW: "How effective was the project in meeting its objective?" What these terms mean from one evaluator to the next or from one program manager to the next or from one context to the next can vary greatly.

*Objective.* Most USAID projects have multiple "objectives" and the background sections of evaluation SOWs don't always shed much light on what is meant by "objective" or distinguish between and activity and its objective. Rather than just stating "objective," write out the specific result or results that the project intended to achieve and that you want investigated *in the evaluation question*. Don't assume the evaluator knows what you mean by the project's objective even if you think you explained it in the background section.

*Effective.* For some, effective might mean, "Are key stakeholders pleased with design of the program and the performance of the implementer." For others, it might mean, "Did a measurable condition change from X to Y over the course of the project." For others, it might mean, "Did the implementer meet output and outcome targets and activity milestones on time." For others, it might mean (akin to impact) "Did the a measurable condition change from X

to something greater than X *as a result of or as caused by* the project with a certain degree of statistical significance.” It might be better in many instances to avoid this term, but if you are going to use it, it will help the evaluator if the SOW communicates how it is being defined in for each evaluation question.

These are just two examples. Other terms that keep me up at night include:

- “Relevance”
- “Efficiency”
- “Impact”
- “Sustainability”
- “Quality”
- “Success”
- “Progress”
- Institutional or organizational or any other type of “capacity”
- “coordination” and “cooperation”

If you are considering using any of these or a hundred other terms like them in your evaluation questions, please make sure that you yourself and your reader know what you mean by them.

**iii. Researchable.** Evaluators who focus on usefulness or utility sometimes retell the joke that goes as follows.

A man walks down a sidewalk at night and sees another man bent over the sidewalk looking down beneath a street lamp as if he is searching for something. The first man asks what he’s doing and the second man says that he dropped his car keys and would the first man help him look for them. They both look for a while and finally the first man says, “I just don’t see anything. Where exactly do you think you dropped them?” The second man says, “Oh, I dropped them way over there”, pointing to a dark, unlit corner of the street. “Then why are you looking here?!?” shouts the first man. The second man replies, “Well, the light is better over here!”

It’s supposed to be a metaphor for the social scientist who studies questions based on the data that he has available not because of the usefulness of the question. It’s a fine story and it makes us feel good about ourselves for asking questions that really matter, unlike some of those silly academics, but I think the corollary to the story should be that the men then go over to the dark corner and then still can’t find the key because neither one has a flashlight, and what’s more, the place where he dropped the key is over a grate that is inaccessible.

It’s great to ask really useful questions, but for the evaluation to be successful, you need to make sure that the useful questions that you are asking are actually researchable with the tools that the evaluator will have. You need to ask if an evaluator will be able to collect sufficient, objective evidence to support an answer to this evaluation question.

Asking a question about the impact of a program on beneficiaries is not researchable if you haven't budgeted for data collection or if you haven't collected baseline data. Asking a question about efficiency of a project is not researchable if you don't have data on project inputs. Asking a question about whether a project was "good" or "successful" are not researchable unless you can define those terms in empirically researchable ways.

#### **4. You don't need a methodology for your evaluation; you need a methodology for *each* evaluation question.**

The way evaluation SOWs are typically structured, there is one section that lists the "evaluation questions" and one section that describes the "evaluation methodology." This makes intuitive sense and would not present much of a problem if there were only one evaluation question being asked. However, it's usually the case that multiple questions are asked in the "evaluation questions" section. For instance, for an evaluation of a local economic development project there might be a question about whether there was a change in a relevant indicator of economic outcomes in the project municipalities, another question about changes in organizational behaviours of municipal governments to produce local economic development plans, another question regarding the various reasons why beneficiaries supported or rejected the project, another question about the likely sustainability of the project, etc.

A problem arises when you have an evaluation questions section that lists multiple evaluation questions, but a methodology section that lists a single methodological approach or even a variety of approaches, but does not indicate which methodologies are suggested for which evaluation questions.

A good statement of work should be clear about what I call the "suggested methodological approach" that is being proposed (more on that phrase in a moment) -- not for the evaluation overall, but for each question that is asked in the evaluation. It is usually the case that you need different methodologies to answer different questions. Let's emphasize that: *you need different evaluation methodologies for different evaluation questions*. A question about the reasons why beneficiaries supported or rejected a project requires a different methodology than a question about municipal organizational capacity.

Too often, I'll see an evaluation with a long list of evaluation questions and then a methodology section that states that the evaluator should conduct interviews, do a survey, and maybe do some focus groups, etc... But for which of the evaluation questions do you expect a survey to be conducted? All of them? Some of them? Which evaluation questions should be answered with focus groups? Tell the evaluation team what you want.

One of the reasons I think this problem arises is that the methodology section is typically not considered in much detail compared to the effort spent on evaluation questions and personnel

qualifications. In my experience, the methodology section works best when it provides a “suggested methodological approach” for each question. By “suggested methodological approach”, I mean that it gives enough detail to give a sense of expectations about data collection methods, analysis methods, and the quality of evidence expected in answering each evaluation question, but not so much detail that the evaluation team does not have flexibility to fill in details and propose some complementary methods. There should be a balance between prescription and flexibility. Usually, I see underspecified evaluation methodology sections rather than over-specified evaluation sections. Providing detailed yet flexible evaluation methodology sections helps by:

- i. It sets expectations for the evaluation team so that they have a sense of what you are expecting in terms of data collection, analysis, and standards of evidence,
- ii. It helps the SOW drafter consider more carefully the drafters own expectations for the evaluation,
- iii. It helps the drafter prepare a more accurate budget estimate.

There are few simple ways to remedy this without changing the structure of the Evaluation Sow. The most simple is to reference each evaluation question when describing the methodology. For instance, write, “In responding to question one, the evaluators shall...”

Another response is to include a simple *evaluation design matrix* in the methodology section. An evaluation design matrix is a table with one row for each evaluation question. The columns in the table address issues such as data collection; data sources, analysis method, and criteria for comparison. Examples are on p. 243 of Morra Imas and Rist, and GAO’s “Designing Evaluations”.

**5. If you expect the evaluator to make judgments about the project being evaluated, they will need criteria for the basis of that judgment.**

Too many evaluations at USAID look like what I would call an “informal review”. In other words, they often consists of an expert in the technical area of the project under review pronouncing judgment on the project and proposing recommendations without (1) presenting an evidence-based argument for those judgments or (2) stating the criteria used for the basis of those judgments. Usually the data collection methods for such studies include document review and interviews with project stakeholders. Now and then, a survey or focus group is thrown in for good measure. In many such evaluations, the data collection methods are inherently limiting in what conclusions can be drawn from them, but the data collection itself is not the really problem, or at least not the biggest problem. The bigger problem is that the presentation of the data and lack of criteria for making evaluative judgments.

For instance in one USAID evaluation, the evaluator wrote, “[the implementer’s] work with service providers within their target service areas was highly professional, effective, and focused on priority health concerns,” without stating what explicit criteria was used for determining “professional”

“effective” or a “focused on priority health concerns” and without presenting the evidence that was collected or reviewed by the evaluator to make this judgment. I’m sure the expert reviewer had seen something or heard something from someone that would be evidence that the service providers were professional and focused on priority health concerns, but the evidence wasn’t presented in an argument to support the judgment. Even when such evaluations do present evidence, judgment requires another step: comparison of the evidence against some criteria – a project target, an external benchmark, a consensus based standard, established best practices, statistical significance. If a service provider spends 20% of their time working on a priority health concerns (however that is defined), does that mean that they are “*focused*” on priority health concerns? What exactly is the criteria for making that determination?

There’s an argument to be made that an expert evaluator is hired for his/her expertise in the technical field of study (and not for expertise in evaluation methodologies) and that it’s OK to rely on their internal, unstated criteria for the value judgments that they make about a program. (See Eliot Eisner’s theory of “connoisseurship” for one of the most articulate arguments for this type of evaluation), but I think this goes against our evaluation policy and its intent. Consider the following bullet points in the evaluation policy regarding the expected features of USAID evaluations:

- “Use of data collection and analytical methods that ensure, to the maximum extent possible, *that if a different, well-qualified evaluator were to undertake the same evaluation, he or she would arrive at the same or similar findings and conclusions,*” and
- “application and use to the maximum extent possible of social-science methods and tools that *reduce the need for evaluator-specific judgments.*”

Evaluator-specific judgments are far too common at USAID. So, if we want the judgments that are not evaluator specific, we need to make sure that the criteria for the basis of the judgment is crystal clear. You can either supply those criteria in the SOW, ask the evaluator to propose criteria, or make the development of the criteria part of the evaluation design process done in collaboration between the evaluation team, USAID evaluation managers, and/or other stakeholders.

## **6. Don’t forget that gathering and analyzing evidence takes resources.**

I think this is fairly self-explanatory. Generally, we under-budget for our evaluations, particularly give the ambitions of our evaluation questions. We also do not adequately allow time for preparation and analysis. Fieldwork/data collection should be one third of your LOE.

If analysis and reporting is not at least one-third of the LOE, maybe the evaluators are not doing enough analysis? And, why do we expect evaluators to have findings and, even worse, draft reports, at the end of their data collection visits to a country. How rigorous do you expect the analysis to be if evaluators are not given time to actually review the data they collected? Providing more resources will not guarantee a better report, but providing too few resources for an evaluation will almost certainly lead to problems in evaluation quality.

## References

Eisner, Elliot. "Educational Connoisseurship and Educational Criticism: Their Form and Functions in Educational Evaluation." *Journal of Aesthetic Education*, Vol.10, No. 3/4, 1976, pp. 135-150

Morra Imas, Linda G., Ray C Rist. 2009. *The Road to Results: Designing and Conducting Effective Development Evaluations*. The World Bank, Washington DC.

U.S. Government Accountability Office. 1991. *Designing Evaluations*. Washington, DC PEMD-10.1.4.