Weakness in evaluation culture may stem from weaknesses in the organization's evaluation system. An evaluation of the World Bank Group's self-evaluation systems finds that there is comprehensive coverage but that systems produce little value to staff, line management, clients, and beneficiaries. Information generated through the systems is not regularly mined for knowledge, learning, or performance management by operational units. There is for instance a widespread blaming and frustration, often around ratings disputes, that limits the appetite to learn from evaluations. Managers and evaluators looking to improve evaluation culture and learning from evaluation systems may want to focus on how well mandatory self-evaluation promotes evaluation utility. For example, is there a way to better target evaluations to decision-makers' key challenges and evidence needs?
Rasmus Heltberg, Lead Evaluation Specialist, Independent Evaluation Group, World Bank Group

Key messages

- The design of self-evaluation systems in large organizations impacts the evaluation culture and evaluation use in those organizations.

- Evaluation culture suffers when evaluation systems do not produce information that is valued by the organization and evaluation is not used for learning.

- Producing evaluations and self-evaluations that will be used and valued by the organization will help to improve evaluation culture.

- To improve learning from evaluation systems, there should be a focus on how the organization uses and values evaluation.

To improve learning from evaluation systems, focus on how the organization uses and values evaluation

An evaluation unit has existed within the World Bank Group (WBG) since 1970. As such, the World Bank has a long history of evaluation and a strong evaluation culture that has grown in size and importance over the years. While initially focusing on project evaluation, the World Bank started self-evaluation of its investments in 1975 and, over time, systems have grown increasingly comprehensive in scope and coverage. Today, self-evaluation lies at the heart of how the WBG and other International Financial Institutions (IFIs) measure their results.

Self-evaluation is done by or for operational units; it builds on monitoring data and feeds into independent evaluation and corporate results measurement, usually via ratings that are independently validated by the independent evaluation offices of the IFIs.

This article reports and reflects on the author’s experience with evaluating the WBG’s self-evaluation systems (World Bank 2016). That evaluation sparked efforts to improve monitoring and reform the WBG’s self-evaluation systems. It has also attracted interest from other IFIs, many of which have self-evaluation systems with similar characteristics and, arguably, pathologies.

Expanding the evaluation framework

Past assessments of self-evaluation systems in IFIs often focused on the degree to which actual practice complied with stated policy and with implicit notions of what an ideal system looked like. Behavioral aspects received limited attention. These assessments often recommended expansion of coverage, assuming that would be a good thing in itself. Assessments didn’t always examine the benefits of doing more evaluation.

The framework for IEG’s evaluation of the WBG’s self-evaluation systems expanded on past assessment frameworks in some ways. It looked at the entire chain from project monitoring and supervision to completion reporting. “validation”
of completion report ratings, and use of the resulting information for portfolio management, overall results measurement, and organizational learning. It focused on mandated self-evaluations but also considered impact evaluation as an example of voluntary self-evaluation.

It analyzed evaluation systems at multiple levels: templates and guidelines; business processes, data streams, roles and responsibilities; and behaviors influenced by norms, values, incentives, and organizational culture. The evaluation also examined how independent validation and evaluation influence self-evaluation processes, behaviors, and incentives.

**Limited use of self-evaluation information**

The stated purposes of self-evaluation give equal weight to accountability, project/program performance management, and organizational learning. But in practice the operation of systems has focused on results reporting for accountability purposes. The systems produce corporate results measures such as the percentage of operations rated satisfactory.

These systems produce little value to staff, line management, clients, and beneficiaries. There has been less attention to developing learning to enhance performance. Information generated through the systems is not regularly mined for knowledge and learning except by IEG. Project and portfolio performance management also doesn’t much rely on evidence from self-evaluation systems. The evaluators encountered widespread blaming and frustration.

We produced a ‘systems map’ to understand the World Bank Group’s evaluation culture and the role of self-evaluation in it. We found that we needed to map the organizational culture in order to understand deeply how actors engage with the evaluation systems and why there was so little learning from them. Figure 1 illustrates some ways in which behaviors, incentives, and organizational culture limit the systems’ utility. There is much focus on project outcome ratings, exacerbated by a competitive organizational culture. Focus on avoiding negative ratings can limit candor. Depth and accuracy in how self-evaluations are written and used suffer. Operational managers face a relentless pressure to keep the pipeline of new projects and disbursements flowing. They have few incentives to invest in project monitoring. Without good data, self-evaluation systems create little value for operational units.

This systems map helps us understand why the perceived value of the knowledge created from self-evaluation systems is low, that too many risks and failures are hidden, safe spaces to learn from failure are missing, and lessons and knowledge are not mined.

The problem is with the mandatory systems, not staff. Staff are genuinely interested in evaluative findings they consider valuable. For example, the World Bank’s impact evaluations tend to be respected and used by staff. Impact evaluations are based on data and methods perceived to be reliable and are carried out in a process that is more congruent with organizational culture. Staff therefore value them more than they do most mandatory self-evaluations.

**Ways forward for more use of self-evaluation**

The issues discussed have been noted in a few individual evaluations and studies (e.g., Weaver 2010; Raimondo 2015; 2016). Reviews of DFID noted heavy reporting burdens on staff and partners that do not seem justified given the use made of the results data, as well as shifting of focus and of incentives toward individual outputs, activities,
and short-term results (OECD DAC 2014; ICAI 2015). But development agencies have by and large not done much to reform how their cultures promote evaluation use.

Arguably, the move toward evaluation systems that are comprehensive, transparent, and mandatory has been fueled by a belief that evaluation is a universal good (that is, a good idea in all circumstances). This belief has led to rules-based, comprehensive evaluation systems. IFIs tend to run their self-evaluation systems using rules-based approaches that spell out what, when, and how operations should be evaluated. Rules that apply across the board help reduce complexity and make things more manageable in large public sector bureaucracies. But rules that apply across the board may also have undesirable unintended consequences such as changes in behavior, large hidden costs, and avoidance of risk and blame (Dahler-Larsen, 2012; Hojlund, 2014; Raimondo, 2016).

What might be a way forward? For starters, consider what evidence decision-makers really need. Is universal reporting based on across-the-board
Evaluation coverage needed? Is there a way in which evaluation could be better targeted to decision-makers’ key challenges and evidence needs? Dahler-Larsen (2012) advocates for a more selective approach to evaluation whereby operations are evaluated when the expected marginal benefit of evaluation exceeds the cost. He proposes a revival of the evaluability assessment:

In other words, to foster a more use-focused evaluation culture, focus on producing evaluations and self-evaluations that will be used and valued by the organization. This would be a marked departure from current practice among IFIs.

References


Rasmus Heltberg is a Lead Evaluation Specialist at the Independent Evaluation Group (IEG) of the World Bank Group. He leads large corporate evaluations, specializing in organizational effectiveness. In his 17 years with the World Bank he has also worked on social protection, social development, adaptation to climate change, and disaster response. He was a member of the core team for the 2014 World Development Report on Risk and Opportunity: Managing Risk for Development. His research has been published in a variety of interdisciplinary development journals, including World Development, Journal of Development Studies, Environment and Development Economics, and Global Environmental Change. Rasmus holds a Ph.D. and M.Sc. in Development Economics from the University of Copenhagen and a M.Sc. in Quantitative Development Economics from the University of Warwick.