Journal of MultiDisciplinary Evaluation
Volume 13, Issue 28, 2017

JMDE
Journal of MultiDisciplinary Evaluation

ISSN 1556-8180
http://www.jmde.com

You Call This Exemplary? Lessons from an Unsung International Evaluation[1]

Douglas Horton[1]
Independent evaluator, University Park, FL

Background: Program evaluators have reached broad agreement on principles for planning and conducting evaluations and on standards for judging their quality. However, many evaluation stakeholders, including key intended users, may judge evaluations on criteria that differ sharply from the professional standards and the criteria we commonly employ in meta-evaluations.

Purpose: This paper highlights the role of academic discipline and epistemic community in judging what is an "exemplary" evaluation, by examining the case of an evaluation that was considered exemplary by professional program evaluators, but methodologically flawed by professionals from different disciplinary traditions.

Setting: The evaluation in question was carried out within the Consultative Group on International Agricultural Research (CGIAR), a research network with a rich tradition of economic impact assessment.

Intervention: NA

Research Design: This is a case study that combines participatory action research and historical analysis.

Data Collection and Analysis: The study is based on the author's personal involvement in the evaluation and on a review of publications and unpublished documents related to the case.

Findings: A team of experienced evaluators applied what are generally considered to be good practices in the program evaluation community. A meta-evaluation led by an experienced program evaluator considered the evaluation to be exemplary. In contrast, within the CGIAR, both the evaluation and the meta-evaluation study were considered to be methodologically flawed and biased. Three lessons related to exemplary evaluation are formulated and elaborated upon:

Lesson 1. Being exemplary is in the eyes of the beholder.
Lesson 2. Epistemic communities are hard nuts to crack.
Lesson 3. You can't win them all.

While the early efforts with program evaluation analyzed here were experienced as failures, a number of subsequent developments have led to greater understanding of diverse evaluation approaches, and some movement toward agreement on what constitutes exemplary evaluation in the CGIAR. Nevertheless, there is still a considerable way to go.

Keywords: *exemplary evaluation; evaluation principles and standards; disciplinary perspectives; epistemic community.*

[1] An early version of this paper was presented at Evaluation 2015 in a session titled, "You call this exemplary? Real life experiences of different evaluation actors in the mosaic of international development evaluation."

Introduction

The 2015 Annual Conference of the American Evaluation Association was dedicated to exemplary evaluation. But who decides which evaluations are exemplary? And based on what criteria and standards?

Based on years of evaluation experience and professional aspirations for the field, program evaluators have reached broad agreement on principles for planning and conducting evaluations and on standards for judging their quality. As members of professional evaluation associations, we often assume that following our guiding principles and program evaluation standards will lead to evaluations that intended users will view as credible and useful. However, it is important to keep in mind that many evaluation stakeholders, including key intended users, may judge evaluations on criteria that differ sharply from our professional standards and the criteria we commonly employ in meta-evaluations.

This paper reflects on an evaluation carried out in 1996 for International Service for National Agricultural Research (the ISNAR evaluation). At the time, I worked for ISNAR as an evaluation specialist. ISNAR was affiliated with the CGIAR, a network of international agricultural research centers established in the 1970s to mobilize modern science to expand the production of basic foods in developing countries. The CGIAR is perhaps best known for its role in the Green Revolution, which, beginning in the 1960s, resulted in expanded use of high-yielding crop varieties, chemical fertilizers, and pesticides, and in rapid growth of food production in many developing regions (Conway, 1997; Cullather, 2010).

The CGIAR has an over-riding "hard-science" culture centered on crop and livestock improvement through the application of modern genetics and related applied sciences. Within the CGIAR, agricultural economists have conducted research studies that have supported program planning, resource allocation, and fundraising. Prominent among these studies have been economic impact assessments of the returns on research investments. In fact, in the international agricultural research community, impact assessment is generally thought of as a sub-discipline of agricultural economics (Alston et al., 1995).

Whereas most CGIAR centers carry out technical research on crops, livestock and natural resources, ISNAR worked to strengthen national agricultural research organizations and systems. To this end, we carried out applied research on agricultural research policies, organization, and management and provided national organizations with advisory services and training. Since our activities focused on capacity development, when it came time to evaluate ISNAR's impacts, we believed that program evaluation frameworks and methods were likely to be more appropriate and useful than traditional economic methods. I still believe we were right. But at the time, our experiment with program evaluation was a rather dismal failure.

Rationale for the ISNAR Evaluation

The CGIAR Secretariat, based at the World Bank, organized External Program and Management Reviews of each center about each five years. In preparation for an external review, each center was expected to assess its own achievements and impacts (CGIAR, 1995b, p. 11). Rather than do this as a "one-off" exercise, in 1996, we decided to conduct the ISNAR assessment as the first phase of a multi-year initiative to develop and test methods for evaluating organizational capacity development.

At the time, impact assessment and evaluation were becoming "hot topics" in the CGIAR, which had entered a period of financial uncertainty and downsizing. In 1995, a high-level donor group requested that the CGIAR improve its governance and "strengthen the assessment of its performance and impact by establishing an independent evaluation function" (CGIAR, 1995a, p. 11). An Impact Assessment and Evaluation Group was established to foster evaluation as an integral part of CGIAR activities (Özgediz, 1998). One of the group's founding members was Eleanor Chelimsky, a highly respected program evaluator who from 1980 to 1994 had directed the Program Evaluation and Methodology Division of the U.S. Government Accountability Office, the largest evaluation unit in the world at the time. Chelimsky had also served as President of the American Evaluation Association. The Impact Assessment and Evaluation Group represented the first systematic attempt to introduce program evaluation into the CGIAR (Horton, 1998). ISNAR's experimentation with program evaluation approaches fit well within this broader initiative.

Design and Conduct of the ISNAR Evaluation[2]

Evaluation objectives

The ISNAR evaluation aimed to achieve three goals:

1. *Methodology development* – to explore the use of methods for assessing organizational performance and impact, which ISNAR and others could use in the future.
2. *Learning and program improvement* – to provide ISNAR with suggestions for improving its activities and performance in the future.
3. *Accountability* – to provide an external review panel with an independent assessment of ISNAR achievements, impacts and constraints over the last five years.

At the time, there was no consensus on the meaning and purpose of impact assessment in the CGIAR or in the broader fields of international development and program evaluation. Whereas Rossi & Freeman (1985, p.185) viewed the purpose of impact assessment as "establishing whether or not an intervention is producing its intended effects," Schalock (1995, p. 13) focused on before-and-after and with-and-without comparisons, seeing impact analysis as determining "whether the program made a difference compared to either no program or an alternative program." A few years later, the Development Assistance Community of the Organization for Economic Cooperation and Development defined impacts broadly as relating to "positive and negative, primary and secondary long-term effects produced by a development intervention, directly or indirectly, intended or unintended' (DAC, 2002, p. 24).

Within the CGIAR, most impact assessments examined the effects of new technologies produced by research programs on crop yields, incomes, or economic surplus – not on the effects of research programs on such ultimate or long-term goals as food security or poverty reduction.

For both substantive and methodological reasons, we decided to focus the ISNAR evaluation on the direct results or effects of ISNAR activities on client organizations, rather than on exploring their possible indirect or long-term effects on such distant goals as food security, poverty reduction or environmental sustainability. This was in line with work at the Evaluation Division of the International Development Research Centre that led to the "Outcome Mapping" approach (Earle et al., 2001; http://www.outcomemapping.ca).

Planning and organizing the ISNAR evaluation

The ISNAR evaluation was collaborative, in that it was led by external evaluators and involved members of ISNAR at key points during the evaluation process. An external team of evaluators took ultimate responsibility for designing the evaluation, collecting much of the data, analyzing and interpreting the results, and reporting on the assessment. ISNAR staff members made substantive contributions to the evaluation design, assembled information on ISNAR activities and outputs, and commented on draft reports. This division of roles and responsibilities reflected ISNAR's strategy of tapping external expertise in organizational assessment while building its own capacity in this area.

To enhance the quality, legitimacy and credibility of the evaluation, and to maximize ISNAR's own learning and capacity development, we contracted a team of international experts in evaluation, organizational development, and agricultural research management, from the International Development Reserch Centre and Universalia Management Group in Canada and from national agricultural research organizations in Brazil and Ethiopia.[3]

At a June 1996 workshop, five external evaluators and ten ISNAR staff members developed an evaluation framework, formulated a list of evaluation questions, defined a set of

[2] This section draws heavily on Horton and Mackay (1998) and Mackay et al. (1998).

[3] The team was led by Ronald Mackay, professor of education and evaluation specialist, Concordia University (Montreal, Canada). The other team members were Jairo Borges-Andrade, organizational psychologist and evaluator, Brazilian Corporation for Agricultural Research and University of Brasilia (Brasilia, Brazil); Seme Debela, plant geneticist and former Director of the Ethiopian Institute of Agricultural Research (Addis Ababa, Ethiopia); Charles Lusthaus, specialist in organizational assessment and development, Universalia Management Group and McGill Univsity (Montreal, Canada); and Terry Smutylo, Director of Evaluation, IDRC (Ottawa, Canada), who was instrumental in the development of Outcome Mapping.

evaluation studies, designed survey instruments, prepared detailed terms of reference for the

Table 1. Evaluation Studies, Study Questions, Data Sources, and Data Collection Methods

Evaluation study	Key study question			Data source	Data collection method
	What are ISNAR's achievements?	What are ISNAR's impacts?	What are ISNAR's constraints?		
	Nature of data collected				
1. Survey of ISNAR stakeholders	Data on perceived sources & quality of ISNAR's achievements	Data on perceived sources & quality of impacts	Data on constraints both external and internal to ISNAR	27 ISNAR stake-holders	Questionnaire, telephone interviews
2. Survey of agriculture leaders	Data on ISNAR input into 10 areas of management	Data relating to 10 areas of management	Data on constraints both external and internal to ISNAR	66 NARO Leaders	Questionnaire, telephone interviews
3. Country case studies	Data on activities, processes and procedures used by ISNAR to achieve results with NAROs	Data on impacts observed in NAROs	Data on constraints ISNAR encounters in working with NAROs	Key informants in Kenya, Morocco and Uruguay	Case studies based on face-to-face interviews during visits to 3 countries
4. Meta-evaluation of ISNAR reviews			Data on potential for reviews as a learning tool	17 ISNAR review reports	Content analyses focusing on utility and quality criteria
5. Review of inventory of ISNAR's outputs	Quantitative data on ISNAR outputs: documents and training events, 1991 - 96	`		ISNAR records on activities and outputs	Use of ISNAR database

external team, developed a budget for the evaluation, and prepared a schedule of the main activities (Mackay et al., 1998, Chapter 2).

Organizational Assessment Model

To guide the evaluation, the evaluation team selected an organizational assessment framework developed by Universalia (Lusthaus et al., 1995), which views an organization's *performance* as being influenced primarily by the external *environment*, organizational *motivation*, and organizational *capacity*. The framework included an inventory of the components or variables making up each of these four organizational dimensions.

In terms of this model, ISNAR's interventions could impact a client organization's external environment, organizational motivation and capacity; and these in turn could influence the organization's performance. However, ISNAR could not have a *direct* impact on a national agricultural research organization's performance, on the subsequent effects on that organization's clients (i.e. on farmers), on farm-level yields, or on broader economic, social or environmental impacts. For this reason, the evaluation studies described below focused on ISNAR's activities, outputs and impacts on the environment,

motivation and capacity of national agricultural research organizations, as opposed to the performance of these organizations. ISNAR's own operating environment, motivation and capacity were of interest insofar as they facilitated or constrained ISNAR's performance.

Data collection and analysis

To answer the question, "*What have been ISNAR's achievements, impacts and constraints over the previous five years?"* (since the previous external), the evaluation team designed a set of five studies that tapped different sources of information and employed different data collection methods:

1. A telephone survey of national agricultural leaders, to obtain structured perceptual data from a sample of ISNAR's clients, on its achievements, impacts, and constraints.
2. A telephone survey of ISNAR stakeholders – for example, officials in donor agencies and development agencies – to gauge its standing in the international community.
3. Case studies of ISNAR's work and results in agricultural research organizations in Uruguay, Kenya, and Morocco
4. A meta-evaluation of ISNAR-commissioned evaluations and reviews, to assess the methods used and obtain evidence of impact
5. Assessment of an inventory of ISNAR's outputs, to provide the evaluation team with structured information on outputs in relation to ISNAR's goals, areas of work, types of activities, and country / regional coverage

Table 1 illustrates the relationships between the key evaluation questions, the data collection methods and sources, and the nature of the data collected.

Involvement of intended users

To encourage utilization of the evaluation results within ISNAR, managers and staff members were involved at key decision points during the evaluation. Involvement of the external review panel was more problematic. In July, the leader of the ISNAR evaluation team met with external review panel members during their initial visit to ISNAR, to discuss the evaluation plan. Later, all communications between the ISNAR evaluation team and the review panel were via the panel's Secretary – an economist at the CGIAR Science Council based at the Food and Agriculture Organization of the United Nations in Rome. Just before the main phase of the review in October 1996 the role of panel Secretary shifted to another economist at the CGIAR Secretariat, based at the World Bank in Washington, DC. After that point, there was no communication between the ISNAR evaluation team and the external review panel.

Limitations of the evaluation approach

One limitation of the evaluation that we were aware of from the start was its heavy reliance on perceptual data. Within the time and budget available, and in the absence of baseline data and experimental treatments, the team attempted to compensate for the lack of "hard evidence" on performance and impacts by using interview techniques that assisted in memory recall and by triangulating results from the five studies and their data sources.

Evaluation Results

Both the type of results presented in the ISNAR evaluation and the form of presentation differed sharply from those found in most CGIAR impact assessments. The leading textbook on agricultural research evaluation focuses on "contributions of research to economic efficiency and the distribution of benefits [which] can be measured as the net present value of research–induced changes in economic surplus" (Alston et al., 1995. p. 502). The most comprehensive and authoritative meta-analysis of agricultural research impact assessments examines 1,886 rates of return on agricultural research investments reported in 292 publications (Alston et al., 2000).

In contrast to these quantitative studies, the ISNAR evaluation was discursive, examining links between ISNAR strategies, activities, products and services delivered, use of ISNAR outputs by national agricultural research organizations, and the changes in organizational motivation and capacities reported by key informants and survey respondents. The evaluation results were summarized in text, a series of tables, and 12 main findings. Results were illustrated with quotations from interviewees and survey respondents.

The ISNAR evaluation panel presented a 35-page summary report with sections on achievements, impacts and constraints faced by ISNAR, supplemented with background documents on the evaluation methodology; results of the surveys of ISNAR stakeholders and agricultural research

leaders in developing countries; case-study reports on impacts in Kenya, Morocco, and Uruguay; and results of a meta-evaluation of ISNAR reviews. The evaluators avoided high-level impact claims, noting that,

> "There is a long and complex series of causal linkages, involving many different actors and a multiplicity of inputs, between ISNAR activities at one extreme and performance [of national agricultural research organizations] at the other. Between the two extremes there is a dynamic chain of achievement–impact relationships, for which reliability and predictability diminish with each successive link beyond the ISNAR intervention. In evaluating ISNAR's impacts it is prudent, therefore, to focus on the primary impacts, i.e. the outcomes and results as identified, observed, or reported in the surveys and case studies" (Mackay et al., 1998, p. 18).

The central findings of the report were summarized as follows:

> "ISNAR has an important role to play in strengthening the management of agricultural research internationally. ISNAR carries out its role with vigor, resulting in a high level of client satisfaction and significant impacts on NARS and their constituent organizations. ISNAR has earned a good reputation for itself in the process" (Mackay et al., 1998 p. 28).

The evaluation report identified areas in which ISNAR work had contributed most to organizational motivation and capacity, vis-à-visother areas where contributions were smaller or even negative. The evaluation concluded that ISNAR had greatest impacts in the areas of agricultural research policy, planning and inter-organizational linkages; intermediate impacts on management of personnel and research-programs; and much smaller impacts on organizational structure and governance, organizational culture and financial management (Mackay et al., 1998, pp. 22-24). It also noted that "the lack of a clear theory of action that explicitly identifies the connections between its goals and objectives, and the actions and resources it employs to achieve these, limits ISNAR's performance as a learning organization." (Mackay et al., 1998, p. 29)

Diverging Assessments of the ISNAR Evaluation

Positive assessment in a meta-evaluation of CGIAR impact assessments

One of the first major activities of the new Impact Assessment and Evaluation Group was to commission a methodological review and synthesis of existing CGIAR ex post impact assessments. An experienced program evaluator, Leslie Cooksy, was contracted to lead the meta-evaluation. Cooksy was a highly respected program evaluator who had worked with Chelimsky at GAO, and served as president of the American Evaluation Association in 2010.

The purpose of the review was to describe the strengths and weaknesses of past evaluation efforts and determine whether the existing body of evidence on center effectiveness could be synthesized to assess overall effects of the CGIAR (Cooksy, 1997a). The documents reviewed were identified through a preliminary assessment of studies sent by Centers in response to a request to submit ex post evaluation studies conducted from 1980 – 1996. Of the 265 documents sent by the Centers, only 87 that presented evidence on center effects were reviewed.

The other 178 studies were excluded because they presented methods for impact evaluation without describing the results of a specific evaluation; described Center activities or made claims of effects without presenting evidence on effects; presented methods or results for research planning (ex-ante evaluation) rather than providing information on the effects of Center activities (ex-post assessment); projected future effects with minimal discussion of current or past trends; or failed to present a clear link between reported effects and Center activities.

The meta-evaluation report noted that a major limitation of the review was the selection process (Cooksy, 1997a, pp. 1-2). Each Center made its own submissions, apparently using different decision rules, even though each responded to the same request letter. This may have led to the observed variety of type and number of documents submitted by different Centers. As a result, Cooksy cautioned that generalizations about the evaluation efforts of the Centers should be made with caution. Despite these selection issues, the meta-evaluation team felt that a number of useful lessons could be drawn from the exercise. Based on the review of the selected 87 documents, the meta-evaluation concluded that

> "The Centers have a wealth of information on their activities, the products of those activities, and the uptake and use of those products. However, the documents reviewed for this project indicate that less information is available on some of the intermediate and longer-range outcomes of the Centers' activities. In addition, the credibility of the linkages between Center activities and the outcomes reported in the documents is often difficult to assess because of insufficient information on methodology and alternative explanations for observed effects. For this reason, a synthesis of evidence on effects across the documents is not defensible" (Cooksy, 1997a, p. 4).

The meta-evaluation team selected 11 studies for a more in-depth review, including the ISNAR evaluation, because they had been explicitly designed to evaluate the effects of center activities (rather than, for example, the effects of a specific new technology) and were broad in scope, in geographic or substantive terms.

The purpose of the in-depth review was to "identify methodological approaches that yielded credible claims of outcomes, and compile those claims that met a high standard of plausibility to learn what could be said with confidence about Center effects in general" (Cooksy, 1997b, p. ES-1). The study did not synthesize the 11 evaluation reports, because: "not all the claims made in the evaluation reports met high standards of plausibility" (Cooksy, 1997b, p. ES-4). The overall conclusion was that

> "after careful review of the 11 documents, we still know very little about the degree to which the CGIAR is achieving its mission of food security and sustainable agriculture in developing countries" (Cooksy, 1997b, Section 4, p. 6).

The report identified 4 promising approaches to developing credible evidence that were modeled by at least one of the evaluations: (1) clear description and rationale for the sampling of cases, respondents, or documents reviewed; (2) synthesis of evidence across multiple sources; (3) disclosure of data gaps and limitations and cautious reporting when faced with severe data gaps and limitations; and (4) use of a logical framework to organize the information presented (Cooksy, 1997b, Section 3, pp. 9-10).

In assessing the ISNAR evaluation, the meta-evaluation report noted that

> "ISNAR's report uses multiple sources of data (surveys, multiple case studies, and syntheses of past evaluation reports) to document its achievements during the period 1991-1996. The strengths of this report include its clear discussion of limitations of the data obtained (such as nonresponse in the surveys) and the use of an external evaluation team. In addition, the evidence from the different sources is presented in the context of ISNAR's activities so that logical connections from activities to products to use can be made" (Cooksy, 1997b, Section 3, p. 5).

According to the meta-evaluation report, while five of the eleven studies made claims about the institutional uptake and use of products, only three of them made *plausible* claims. The ISNAR evaluation was considered the most successful in supporting its claims:

> "ISNAR is most successful in supporting a range of claims… Not only does it support claims of public awareness of its products, it is the only one of the three centers that presents plausible evidence that its recommendations (specifically its approach to planning and priority setting) have been adopted. Note that ISNAR also measured customer satisfaction, which, while not necessarily linked to effectiveness, can be an indicator of the use of center products" (Cooksy, 1997b, p. 4-1).

Negative assessment by the external review panel

The external review panel had little use for the ISNAR evaluation. While noting several strengths of the evaluation,[4] it identified key weaknesses in

[4] The review report noted the following virtues of the ISNR evaluation: The evaluation report contained valuable information on ISNAR outputs and on the perceptions of NARS leaders and stakeholders regarding ISNAR's activities and reputation; the externality of the study team members contributed to the objectivity of the report; the data-gathering effort was based on an organizational assessment framework; the interviews were conducted professionally; and the surveys provided useful information on client and stakeholder views on ISNAR's work.

the methodological approach related to the measurement of the ISNAR-related interventions, the measurement of changes in organizational performance, and attribution of changes in performance to the ISNAR intervention. The performance indicators used in the surveys were viewed as "necessarily all subjective, given the methodology." Moreover, "the methodology used did not allow control of factors other than the ISNAR intervention influencing [national agricultural research organizations'] performance." (TAC Secretariat 1997, p. 18). The panel concluded that the study was "an experiment in assessing organizational performance and one that should be used as a stepping stone for building, in due course, more rigorous tools for institutional evaluation." The panel urged ISNAR "to insist on greater methodological rigor" in future work in this area (TAC Secretariat 1997, p. 18). During the main phase of the review, the panel secretary mentioned that he suspected the evaluation data "had been cooked."

Negative assessment of the CGIAR meta-evaluation

Jock Anderson, a senior agricultural economist in the World Bank's Operations Evaluation Department (now Independent Evaluation Group) with extensive experience evaluating impacts of international agricultural research, was invited to comment on the Cooksy meta-evaluation study at the annual meeting of the CGIAR in October 1997. Anderson, who had led a comprehensive assessment of the CGIAR's achievements and impacts in the mid-1980s (Anderson et al., 1988), opened his comments with a clear statement of concern: "It is good for evaluators to be 'lean and mean,' but when they are only mean, one can be justifiably concerned" (Anderson, 1997, p.1). He expressed special concern for the sampling procedures used in the meta-evaluation, noting that few of the background studies and none of the main reports prepared for the mid-1980s system-wide impact assessment were included in the meta-evaluation.

> "It is apparent that the process engaged in did not meet the test of unbiased sampling (in this case, of the global literature) that is recommended as good practice in judging the acceptability of impact evidence… One aspect of the non-inclusion of the mid-1980s Impact Study that is especially galling, at least to me, relates not to rejection but to non-consideration at all. A constant theme in Reports 1 and 2 is a quest for meta-evaluation across the system within a consistent framework … The previous Impact Study [Anderson, Herdt and Scobie, 1988] did this explicitly across the whole System" (Anderson, 1997, p. 3).

Concerning the in-depth analysis of 11 impact studies, Anderson was highly critical:

> "The report is long on statements about plausibility (or, more precisely, lack of it!) in the appreciation of Center documents on impact. The subjectivity inherent in these types of comments is great, as it is necessarily now also in my mentioning that I find most of the analysis in this report to be quite implausible and thus not very useful in itself. If such critical works help to give us cause for pause, and to think more carefully and cogently about the persuasiveness of evidence of different types, perhaps some good will have been achieved, but that proposition remains to be tested." (Anderson, 1997, p. 4)

In light of this highly critical assessment, the results of the meta-evaluation were never formally issued by the CGIAR or published elsewhere.[5]

In the same presentation, Anderson also criticized another evaluation study, on factors affecting adoption and impacts of CGIAR innovations, led by the well-known program evaluator Lee Sechrest (1991 president of the American Evaluation Association):

> "I would hope that we might hear from some of the Centers here what they think of this work, which does not impress me as having significantly advanced the research agenda over that in many earlier Center studies of such phenomena, and has none of the broad, fresh thinking outlined in other contemporary contemplations of such matters … Let me make clear that I see nothing but good coming from fresh approaches to impact investigation, making greater use of procedures developed in the evaluation community, but let's do our homework and get it right before squandering too

[5] The methods employed in the meta-evaluation were published in the American Journal of Evaluation (Cooksy and Caracelli, 2005).

> many scarce resources in the name of better evaluation" (Anderson, 1997, pp. 4-5).

Thus, the first 2 major studies carried out by experienced and respected program evaluators were roundly and publically criticized – essentially given failing marks – by one of leading authorities on impact assessment in international agricultural research.

Subsequent developments

After its initial trials and tribulations, the Impact Assessment and Evaluation Group was disbanded, bringing to a close this first system-wide experiment with program evaluation. In its place, a Standing Panel on Impact Assessment was established to "provide CGIAR members with credible information about the impact of past CGIAR research centre investments and to enhance and systematize more the quality of ongoing CGIAR centre ex post impact assessment efforts" (Kelly et al. 2008, p. 202). A parallel Standing Panel on Monitoring and Evaluation was also established mainly to coordinated external reviews of centers. Since its establishment, the impact assessment panel has been led by agricultural economists who are considered leaders in the field of impact assessment. Over the years, this panel has supported methodological development and organized numerous impact assessment studies, results of which have been published in books and professional journal articles. In contrast, the monitoring and evaluation panel was chaired by research managers without professional standing in the field of program evaluation, and it had no similar record of methodological advances or evaluation studies. In a recent reorganization of the CGIAR, the monitoring and evaluation panel was disbanded, but the impact assessment panel continues to operate as a key component of the Independent Science and Partnership Council of the CGIAR (http://impact.cgiar.org).

One of the first activities of the impact assessment panel was to commission Prabhu Pingali, Director of the Economics Program at the International Maize and Wheat Improvement Center to prepare a study of "milestones in impact assessment research" (Pingali, 2001). Pingali's study covered only research on the impacts of new technologies and agronomic management practices, excluding the contributions of CGIAR scientists to impact assessment research related to policy research and advice and capacity strengthening. The rationale for this exclusion was that these areas had received inadequate attention by evaluators (Pingali, 2001, p.1). Pingali's review concluded that

> "Over the past three decades, CGIAR economists have been actively involved in assessing the adoption and impact of MVs [modern varieties] and other technologies developed by the CGIAR Centers. Impact assessment at the CGIAR has been recognized for its substantive and methodological contributions by the economics profession as well as by the donor community that invests in the CGIAR centers… Numerous high-quality publications in international journals, as well as numerous awards received by CGIAR researchers, attest to the high quality of the [impact assessment] research conducted at the centers" (Pingali, 2001, pp. 12-13).

Neither the ISNAR evaluation nor the meta-evaluation of CGIAR impact assessments was mentioned in this review.

Lessons

The ISNAR evaluation and the meta-evaluation of CGIAR impact assessments suggest three broad lessons concerning exemplary evaluation.

Lesson 1: Being exemplary is in the eyes of the beholder

There are no universal criteria or standards for judging an evaluation. Program evaluators have formulated useful principles and norms to guide our work. But at the end of the day, stakeholders may look at our evaluations through very different lenses and judge them on the basis of criteria and standards that differ sharply from those of the program evaluation community.

In an assessment of the effectiveness of knowledge systems for sustainable development, Cash et al. (2003) note that scientific information is likely to influence debates to the extent that the information is perceived by relevant stakeholders to be credible, salient, and legitimate. In this context, *credibility* involves the "scientific adequacy of the technical evidence and arguments;" *salience* refers to the "relevance of the assessment to the needs of decision makers;" and *legitimacy* "reflects the perception that the

production of information has been … unbiased in its conduct" (Cash et al., 2003, p. 8087). The same authors emphasize the importance of "boundary work" to facilitate communication, translation, and mediation across the boundaries between experts and decision makers.

These principles can be usefully extrapolated to the field of evaluation. In both the ISNAR evaluation and the meta-evaluation of CGIAR impact assessments, boundary work was inadequate. Relevant stakeholders questioned the scientific adequacy of the methods employed and felt the studies were biased. One implication is that to encourage understanding and a favorable perception of an evaluation, evaluators should engage key stakeholders to the extent possible throughout the evaluation process. Evaluators will recognize this as a basic principle of utilization-focused evaluation (Patton, 2008).

Lesson 2: Epistemic communities are hard nuts to crack

In a classic article, Haas (1992, p. 3) defines an epistemic community as "a network of professionals with recognized expertise and competence in a particular domain and an authoritative claim to policy-relevant knowledge within that domain or issue area." In the field of international agricultural research, agricultural economists form such a community. They fought long and hard to gain acceptance and authority in policy discussions in the hard-science culture of the CGIAR, and have proven highly effective in that sphere. Conducting impact assessments has been central to gaining acceptance and authority. Ex-ante assessments have contributed to priority setting and resource allocation within the CGIAR, and ex-post assessments that have shown high returns on investments in agricultural research, have helping justify funding for CGIAR centers. Over the years, agricultural economists have risen to many of the senior-most positions in agricultural research management and governance. The location of the CGIAR Secretariat (more recently Fund Council) at the World Bank has further cemented the power and prominence of economists in the CGIAR. As can be imagined, they are not about to cede ground in the evaluation field to "a bunch of social scientists."

Lesson 3: You can't win them all

The ISNAR evaluation was by no means perfect. But it is unlikely that any attempt to assess ISNAR's impacts would have been viewed as "exemplary" in the CGIAR. One reason was that impact assessment was considered a branch of agricultural economics. So even though the methods employed by agricultural economists were inadequate for evaluating capacity development and institutional strengthening, an impact assessment that employed other methods – such as those associated with program evaluation – was bound to fail on methodological grounds. It simply would not be considered a "real" or "legitimate" impact assessment.

A second reason was that ISNAR itself was always controversial. In a system where most centers were dedicated to scientific research and the generation of new farming technologies, ISNAR was a poor fit. Throughout its existence, controversy surrounded ISNAR's mission and its role in international agricultural research. These controversies only ended when ISNAR was closed in 2004. Since that date, the CGIAR has had no concerted effort to strengthen national agricultural research or agricultural innovation systems in the global South.

Subsequent Developments

While these early efforts produced few positive results in the short term, over time, numerous developments have led to greater convergence on what is considered "exemplary" evaluation in the CGIAR. Several program evaluators, including Frans Leeuw, John Mayne, Elliott Stern, and Leslie Cooksy have been involved in evaluation activities, at the level of CGIAR programs and also at the system governance level. An "Institutional Learning and Change Initiative" established in 2004 spearheaded the use of learning-orientated evaluation in the CGIAR for a decade.[6] In 2008, the CGIAR issued guidelines for ex post impact assessment that included several program evaluation approaches (Walker et al., 2008). Use of theory of change has been introduced into planning and evaluation procedures for CGIAR global programs. Most important of all, in 2012 at the request of its major donors, the CGIAR prepared an evaluation policy and established an Independent Evaluation Arrangement (http://iea.cgiar.org) modeled on similar evaluation groups in the Food and Agriculture Organization of the United Nations, the World Bank and other international agencies.

[6] Impetus for establishing "Better Evaluation" (http://betterevaluation.org) emerged from an ILAC-sponsored workshop in 2008.

Conclusions

This paper has reflected on early experiences with program evaluation in an international network of research centers in the 1990s. It illustrates how different stakeholder groups representing different disciplines and epistemic communities can have sharply differing perceptions of the quality and usefulness of an evaluation. What one group considers exemplary, other groups may consider suspect, methodologically weak, or biased. While the early efforts with program evaluation analyzed here were experienced as failures, a number of subsequent developments have led to greater understanding of diverse evaluation approaches, and some movement toward agreement on what constitutes exemplary evaluation in the CGIAR. Nevertheless, there is still a considerable way to go.

References

Alston, J., Norton, G., & Pardey, P. (1995). *Science under scarcity: Principles and practice for agricultural research evaluation and priority setting*. Ithaca, NY: Cornell University Press.

Alston, J., Chan-Kang, C., Marra, M., Pardey, P., & Wyatt, T. (2000). *A meta-analysis of rates of return to agricultural R&D: Ex pede herculem?* Research Report 113. Washington, D.C.: International Food Policy Research Institute.

Anderson, J., Herdt, R. & Scobie, G. (1988) *Science and food: The CGIAR and its partners*. Washington, D. C.: The World Bank.

Anderson, J. (1997). *Draft observations on the impact presentations*. Notes for a presentation to the CGIAR International Centers Week, October 31, 1997. http://library.cgiar.org/bitstream/handle/10947/480/cg9710d.pdf?sequence=1

Cash, D., Clark, W., Alcock, F., Dickson, N., Eckley, N., Guston, D., Jager, J. & Mitchell, R. (2003). Knowledge systems for sustainable development. *PNAS* 100 (14), 8086-8091.

CGIAR Secretariat. (1995a). *Renewal of the CGIAR: Sustainable agriculture for food security in developing countries*. Washington, DC: Consultative Group on International Agricultural Research.

CGIAR Secretariat. (1995b). *Review processes in the CGIAR: Terms of reference and guidelines for external program and management review of CGIAR centers*. (CGIAR document No. MTM/95/11). Washington, DC: Consultative Group on International Agricultural Research.

Conway, G. (1997). *The doubly green revolution: Food for all in the 21st century*. London: Penguin Books.

Cooksy, L. & Caracelli, V. (2005). Quality, context, and use: Issues in achieving the goals of metaevaluation. *American Journal of Evaluation* 26 (1), 31-42.

Cooksy, L. (1997a). *CGIAR methodological review and synthesis of existing ex post impact assessments. Report 1: A review of documents reporting effects of international agricultural research centers*. Rome: Consultative Group on International Agricultural Research, Impact Assessment and Evaluation Group.

Cooksy, L. (1997b). *CGIAR methodological review and synthesis of existing ex post impact assessments. Report 2: Analysis of comprehensive ex post studies of impacts of international agricultural research centers*. Rome: Consultative Group on International Agricultural research, Impact Assessment and Evaluation Group.

Cullather, N. (2010). *The hungry world: America's cold war battle against poverty in Asia*. Cambridge: Harvard University Press.

DAC. (2002). *Glossary of key terms in evaluation and results-based management*. Organization for Economic Cooperation and Development (OECD), Development Assistance Committee. http://www.oecd.org/dac/2754804.pdf

Earle, S., Carden, F. & Smutylo, T. (2001). *Outcome mapping: Building learning and reflection into development programs*. Ottawa: International Development Research Centre. http://www.outcomemapping.ca/download/OM_English_final.pdf

Haas, P. (1992). Introduction: Epistemic communities and international policy coordination. *International Organization*. 46 (1), 1-35.

Horton, D. (1998). Disciplinary roots and branches of evaluation: Some lessons from agricultural research. *Knowledge and Policy*. 10 (4), 31-66.

Horton, D. and Mackay, R. (1998). Assessing the organizational impact of development cooperation: A case from agricultural R&D. *Canadian Journal of Program Evaluation* 13 (2), 1-28.

Kelley, T., Ryan, J. & Gregersen, H. (2008). Enhancing ex post impact assessment of agricultural research: the CGIAR experience. *Research Evaluation* 17 (3), 201–212.

Lusthaus, C., Anderson, G., & Murphy, E. (1995). *Institutional assessment: A framework for*

strengthening organizational capacity for IDRC's research partners. Ottawa: International Development Research Center.

Mackay, R., Debela, S., Smutylo, T., Borges-Andrade, J., & Lusthaus, C. (1998). *ISNAR's achievements, impacts, and constraints: An assessment of organizational performance and institutional impact*. The Hague: International Service for National Agricultural Research.

Patton, M. (2008). *Utilization-focused evaluation*. Los Angeles: Sage Publications.

Pingali, P. (2001). *Milestones in impact assessment research in the CGIAR, 1970-1999*. Rome: Standing Panel on Impact Assessment, Technical Advisory Committee of the Consultative Group on International Agricultural Research.

Özgediz, S. (1999). Evaluating research institutions: Lessons from the CGIAR. *Knowledge and Policy*. 11 (4), 97-113.

Rossi, P. & Freeman, H. 1985. *Evaluation: A systematic approach*. Newbury Park, CA: Sage Publications.

Schalock, R. 1995. *Outcome-based evaluation*. New York, NY: Plenum Press.

TAC Secretariat. (1997). *Report of the third external programme and management review of the International Service for National Agricultural Research*. Document No. SDR/TAC:IAR/96/23. Rome: Technical Advisory Committee of CGIAR.

Walker, T., Maredia, M., Kelley, T., La Rovere, R., Templeton, D., Thiele, G., & Douthwaite B. (2008). *Strategic guidance for ex post impact assessment of agricultural research*. Rome: Standing Panel on Impact Assessment, CGIAR Science Council Secretariat.