IEA WORKSHOP ON
EVALUATING QUALITY OF SCIENCE

Report

Rome 10-11 December 2015

Independent
Evaluation
Arrangement
CGIAR

1. INTRODUCTION

For a research for development organization such as CGIAR, the quality of science (QoS) in a program is an important determinant of its potential effectiveness. In the external evaluations of CGIAR research programs (CRPs), QoS is an important component among other criteria of relevance, effectiveness, impact, management efficiency and sustainability (that are in line with the OECD-DAC evaluation criteria)[1] to reflect the particular mandate of the CGIAR. In the past, the External Program and Management Reviews (EPMR) of Centers covered QoS, though there was no consistent approach regarding the dimensions of science quality to be assessed. Scientific publications were commonly used as a hallmark of scientific output, and assessment focused on bibliometric analysis and assessment of research staff.

Since the 2010 CGIAR Reform, evaluations have shifted their focus from Centers to CRPs. With the establishment of the Independent Evaluation Arrangement (IEA), a broad framework for evaluating QoS was adopted that involves assessing inputs, processes and outputs. By early 2016, IEA will have completed 10 CRP evaluations that have broadly followed this framework. Through its work, IEA aims at providing consistent and comparable evaluative information on science quality in the CRPs and contributing to and enhancing the conceptualization of QoS in the R4D framework at CGIAR centers and programs.

Recently, in the CGIAR 2013 Mid-Term Review (MTR), assessing and ensuring high quality QoS were raised as important issues for CGIAR to address. The MTR highlighted the role of the Independent Science and Partnership Council (ISPC) in advising the CGIAR System on issues of science quality and promoting high QoS, while also highlighting the Consortium Office's (CO) role in monitoring QoS at the CRP level.

The assessment of QoS in program proposal ex-ante appraisal (ISPC), program monitoring (CO) and program evaluation (IEA) are somewhat different, as are the parameters, which are assessed. However, as the processes, issues, and areas of focus are closely linked, and all functions are intended to inform the programs, donors and other stakeholders, agreement in definitions is needed and consistency in the criteria and assessments is desirable.

This workshop report consists of sections that deal with (i) the purpose of the workshop; (ii) summary of presentations and discussion from the workshop sessions; (iii) reflection of the main issues addressed during the discussions; and (iv) conclusions for the future.

2. PURPOSE OF WORKSHOP

The IEA organized a 1.5 day workshop to bring together a small group of stakeholders from within and outside the IEA to discuss approaches and experiences relating to QoS assessment in evaluation of CRPs.

The overall purpose of the workshop was to consolidate and strengthen the IEA's approach to evaluating QoS in the CGIAR, focusing on the evaluation framework, evidence used, interpretation of findings, as well as lessons for users of evaluation. The workshop was also seen as an opportunity to

[1] See http://www.iea.cgiar.org/sites/default/files/Standards.pdf

explore the scope for achieving a common understanding and definition of OoS and how to link it to other aspects of performance in appraisal, monitoring and evaluation.

The stated workshop objectives were:

- assess the implementation of the IEA QoS evaluation framework in the completed CRP evaluations;
- critically review strengths and challenges for collecting data and generating evidence for QoS assessment, and the subsequent interpretation in evaluations;
- learn from reflection on experience elsewhere within and outside CGIAR;
- enhance a common understanding of QoS assessment and appraisal in CGIAR, and in particular CRPs, across the three units (IEA, ISPC, CO);
- strengthen QoS evaluation in the CGIAR, as part of a broader set of evaluation criteria, for improving its usefulness;
- draw broader lessons for the CGIAR with the aim of preparing guidelines for evaluation of QoS.

A total of 16 participants attended the workshop. In addition to IEA evaluation staff, participants included representatives from other CGIAR System units (ISPC & ISPC Secretariat, Consortium Office); and IEA evaluation team members. External participants came from IDRC, GIZ (representing a donor agency) and one CRP (also representing research establishment outside the CGIAR). For a complete list of participants, see Annex A.

3. SESSION OVERVIEW

The workshop was structured into five sessions:

- Session 1 set the stage and outlined the context in which IEA developed its framework for assessing QoS.
- Session 2 focused on reflecting on the experiences to date in implementing the evaluation framework.
- Session 3 focused on external perspectives and experiences, with presentations from ISPC and IDRC.
- For Session 4, two working groups were formed to discuss potential improvements to the framework and propose potential indicators and how they should be used for assessing QoS in CRPs.
- Session 5 was a wrap up session, providing an opportunity to reflect on the discussions during the workshop and potential ways forward.

Session 1: IEA Quality of Science evaluation framework

Chair: Rachel Sauvinet-Bedouin, IEA Head
Presentation: Sirkka Immonen, IEA
Group discussion

Currently, CGIAR does not have a common understanding or definition of QoS. The past Center EPMRs were conducted by peers using sources of evidence and methods of analysis similar to those in current IEA evaluations, but often restricted to bibliometric and researcher analysis. A broad aspect of Center performance assessed was "the quality and relevance of science", and there was a lack of consistency and comparability across the EPMRs. For some time (2005-2009), the monitoring of QoS was based on a performance measurement system, which mainly focused on the volume of publications. While indicators (such as number of publications) are proxies of some aspects of performance, evaluations can actually provide contextualized and holistic analysis of performance through interpretation of quantitative and qualitative information.

The QoS framework that IEA introduced for CRP evaluations is three-dimensional and looks at (1) inputs, (2) management processes and (3) outputs (see Table 1).

The challenges identified by the IEA regarding the overall framework include:

- "losing the forest for the trees" (long list of items to potentially include);
- not linking the essential QoS issues to the rest of the evaluation and to other evaluation criteria;
- peers have problems assessing their peers (they may be overly sympathetic).

Session 2: Learning from IEA experience

Chair: Sirkka Immonen, IEA
Presentation: Assessment of Quality of Science Overview from CRP evaluations, Sophie Zimm, IEA
Presentation: Lessons learned from CRP evaluation synthesis, Christian Roth, CSIRO (current lead consultant on the IEA Synthesis of Completed CRP evaluations)
Panel discussion: Former evaluation team members and managers

To date, IEA has completed five CRP evaluations and five are nearing finalization. In preparation of this workshop, a comparative analysis was carried out of how QoS was assessed in these evaluations. The CRP evaluations have applied a mix of methods to assess QoS at the three different levels; inputs, management processes, and outputs. The QoS criterion has evolved since the first evaluation conducted by IEA. This evaluation criterion has currently the most elaborated framework for assessment. Unlike with most other evaluation criteria, quantitative data are available on different indicators for QoS, and the framework has relied on those indicators to a considerable extent. Evaluation evidence for assessment is received from data on researchers, publications and other outputs and through interviews, field visits, observations, expert knowledge, documentation review, as well as review of evaluative information such as other evaluations of the CRP and ISPC reviews and strategic studies.

Across the evaluations, three major issues are emerging:

- First, the role and responsibility of Center vs CRP, where Centers are responsible for QoS of their contribution to the CRP, HR management, infrastructure, and research leadership, and CRPs are responsible for the implementation of the CRP;
- Second is the scope of QoS assessment where, for example, outputs such as publications have largely reflected research done before CRP, inclusion of non-peer reviewed publications that may have other purpose than scientific communication, assessing other non-publication outputs such as breeding materials, modeling portals, tools etc.;
- A third major issue is the extent to which issues related to research process and knowledge management are included in assessment of QoS.

Table 1 summarizes the assessment framework for QoS used in the CRP evaluations.

Table 1. IEA assessment framework for QoS

WHAT WAS COVERED?	HOW WAS IT ASSESSED?
INPUTS *Human resource quality: leadership, team profiles, Research design* *Data management, Facilities, resources, research support, Research partners*	• H index analysis • Analysis of researcher qualifications/time • Interaction with research leaders and teams • Interviews • Evaluative studies
MANAGEMENT PROCESSES *Overall management of QoS, Processes in place for enhancing and assuring QoS, Incentives*	• Researcher survey • Document review (review of center policies, etc) • Observations, interviews
OUTPUTS *Research publications, Non-publication outputs, Research processes (state-of-the-art and efficiency in producing outputs)*	• Qualitative review/peer review style of randomly selected publications • Bibliometric analysis (journal frequency, citation, affiliations) • Evaluative studies like Elsevier study or Stripe reviews • Peer assessment of non-publication outputs • Peer assessment of research processes for generating outputs

Source: IEA.

The relative emphasis of evaluation criteria coverage varied in the first five completed IEA CRP evaluations as shown in Table 2.

Table 2. Evaluation criteria coverage in five CRP evaluations fully completed

Evaluation criteria	AAS	FTA	MAIZE	PIM	WHEAT
Inputs					
Human resource quality	++	+	++	++	++
Research design	++	(++)	++	+	+
Data management	+	(++)	+	+	+
Facilities			++		+
Research partnerships	+	(+)	++	+	
Management					
Overall QoS management	++	(++)	(+)	++	+
QA processes	+	+	(++)	++	+
Incentives	+		++	+	+
Outputs					
Research publications	++	++	++	++	++
Non-publication outputs			++		++
Research processes	++	+	(++)		+

Source: IEA. () denotes indicators covered solely or partially under other evaluation criteria.

Bibliometric analysis was systematically used in all evaluations to assess publication outputs, and in most cases included journal frequency analysis, impact factor of journals (JCR impact factors), citation analysis (google scholar) and affiliations analysis. However, the resulting quantitative data were used to different extent and interpreted differently in the evaluations. In the case of citations, for example, the number of journal articles varied significantly. For example, while the AAS evaluation included only 27 journal articles in the sample, the GRISP evaluation included a total of 769 journal articles.

Also H index analysis was conducted in all evaluations but one (AAS). The underlying sample of researchers varied, however, significantly. While the evaluation of Water, Land and Ecosystem included all project leaders (total of 119), Livestock and Fish (L&F) evaluation only looked at Flagship leaders and Value Chain leaders (total of 19). In interpretation of the data, evaluations variably referred to averages and the distribution of H indexes.

The main lessons and conclusions from the two presentations were as follows:

- With the IEA's QoS evaluation framework having been put in place, QoS assessment has evolved considerably towards better comprehensiveness since the first IEA evaluation;
- QoS is the criterion with the most formal framework for assessment;
- QoS is the criterion where quantitative data are available which is not the case with other criteria;
- There are big differences in the interpretation of quantitative QoS data and in combining them with qualitative information in triangulation;
- In presentation and interpretation of quantitative data averages hide variation and evaluation should assess the distribution and the size of the extremes;
- There are questions around scope of QoS assessment in terms of: representativeness of the evidence of CRP performance rather than pre-CRP research; inclusion of activities and outputs related to research that may need to be assessed from QoS perspective (e.g. non-peer reviewed publications) and coverage of non-publication outputs;
- Most aspects of QoS require assessment of Center products, practices and capacities in order for the evaluation to make conclusions on the CRP performance.

Other lessons and conclusions specific to the dimensions of the framework are the following

Inputs

- The IEA QoS framework broadly captures quality of main inputs; though evaluations used different combinations of evidence;
- There are examples of how bibliographic analysis can be used to assess quality of scientists, but leadership attributes in terms of the leaders' ability to drive excellence and integrate research at the CRP level remain challenging to assess and benchmark in any meaningful way and assessments could be sharper, in terms of definite conclusions;
- Limited systematic analysis of research design: research design is ambiguous because of overlap with other evaluation criteria and assessment of the quality attributes is qualitative in the absence of meaningful indicators. Yet, research design is a CRP responsibility and highly relevant aspect of QoS to assess as it influences likely effectiveness and value added of the Program.

Management

- Evaluations draw mainly on the results of the staff survey;
- Staff surveys provide potentially useful and systematic information across evaluations although in limited aspects of science quality management, but variations in survey questions have reduced comparability, and integration in triangulation with other sources of evidence has been limited;
- Interpretation of evidence and drawing conclusions have been hampered by the lack of full understanding and assessment of the Centers' role in QoS management at the CRP level where Centers' and CRP's roles are intrinsically linked.

Outputs

- Bibliometric data can be interpreted very differently, as there are no universally applicable benchmarks or shared benchmarks or standards across CGIAR evaluations;
- Errors and inconsistencies in records available to the teams, and the fact that CRPs do not have sufficient "branding" for their published outputs to be identifiable in Web databases, make attribution of outputs and subsequently their evaluation problematic;
- Need to improve and define the distinction between legacy and CRP initiated outputs;
- Non-academic publications may need more formalized evaluation in the context of research outcome/impact.

It was highlighted that 'systems science' and transdisciplinary research require tailoring the way in which science quality assessment is applied in the evaluation in relation to other criteria such as relevance. This implies developing approaches and indicators to evaluate progress on systems science, integration and inter/transdisciplinarity. Doing so would require a prior definition of 'systems' and the scope of systems research in the CGIAR.

In the discussion in Sessions 1 and 2, participants brought up the potential confusion following from commonly using the two terms, *science* and *research*, interchangeably. While quality of the science (scientific knowledge) produced by CGIAR research is an essential topic in evaluation, quality of research was acknowledged to be broader. Research process encompasses also the process of producing the knowledge, including the modalities of implementing research, partners conducting research and other aspects of the process. Research process also links science and relevance, when choice of research

topics is considered as part of the process. It was agreed that there is a need to reach common understanding of how the terms science and research are used. Furthermore, participants highlighted the importance of relating quality of science and the research process to other aspects of overall programmatic assessment. These issues were discussed in the subsequent sessions.

Session 3: Learning from other experiences on QoS assessment

Chair: Tim Kelley, ISPC
Presentation: Assessing science quality ex-ante in CRP proposals, Maggie Gill, ISPC
Presentation: IDRC's Research Quality Plus Framework - Design and initial lessons from application, Tricia Wind, IDRC
Group discussion

PART 1 (Maggie Gill, Chair, ISPC Council)

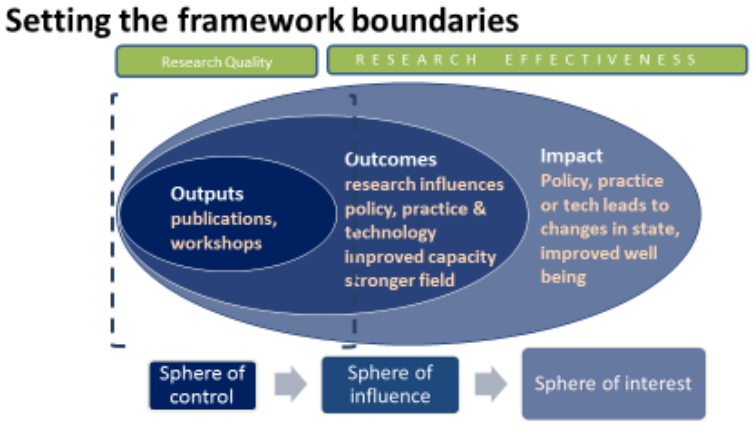
In CGIAR, ISPC assesses QoS at proposal stages of the CRPs. The criteria used by the ISPC with regard to QoS assessment are:

Overall CRP level	• Rigor and credibility of the scientific arguments underpinning the rational for the pre-proposal • Track record of the Leadership Team, including FP leaders (recruitment criteria if leaders not in place)
Flagship level	• Novelty and soundness of the research being proposed • Track record of the FP leadership and team, assessed on the basis of what was achieved in the previous CRP portfolio (publications and demonstration of commitment to quality, peer review mechanisms, etc.) • Lessons learned; evidence of building on previous work (1st round of CRPs); e.g. how things have changed or even been dropped on the basis of past learning

Currently, ISPC is focusing on addressing two key issues: how to combine different ratings of QoS for an overall CRP score; as well as how to assess the leadership of the CRP with respect to the quality of the research at the project level within CRPs, given that Centers retain overall responsibility for QoS in research undertaken by their scientists.

PART 2 (Tricia Wind, Senior Program Officer, IDRC)

The IDRC recently adopted a new framework for assessing research quality called "Research Quality Plus". It focuses on research outputs, as shown in the diagram illustrating the boundaries of the framework.



Source: Tricia Wind, IDRC.

In the IDRC framework, research quality has four different dimensions:

1. Research integrity – scientific merit and methodological rigor of approach and execution of research
2. Research legitimacy - inclusiveness, engagement of local knowledge, gender responsiveness, and addressing potentially negative consequences
3. Research importance – originality and relevance
4. Positioning for Use - knowledge accessibility and sharing; timeliness and actionability

Scientific merit is assessed under research integrity through main research outputs produced. In their analysis of research quality IDRC considers a list of influencing factors such risks in data, research and political environment, research capacity and maturity of the research field.

Session 4: Working groups

- *Two breakout groups*
- *Reporting back*
- *Group discussion*

The two working groups were asked to review the IEA framework and methods most typically used in CRP evaluations (see Table 1). The key questions were:

- *Are the topics clearly associated with QoS; is something missing?*
- *Can and should the topics be prioritized; what are the essential ones for high quality of science?*
- *How can the evidence be used (triangulation, weights, benchmarking, critical thresholds, caveats with individual indicators) for deriving at credible and useful conclusions?*
- *Is assessment of this dimension critically dependent on or linked to assessment of other dimensions of QoS or overall performance and how?*

Both groups presented their thoughts on the different dimensions, including aspects and possible indicators to consider. The participants agreed that the QoS evaluation framework should cover items

under the three dimensions used but suggested revision of the dimensions as: (i) resources and their management (reflecting inputs); (ii) research process (reflecting research management); and (iii) outputs. The items to include under each dimension and information/indicators were discussed as illustrated in Table 3. The complexity of the table partly reflects the multiple aspects of QoS within each dimension and the fact that the groups did not have time to discuss priorities within this long list as to what would constitute a minimum set of indicators/information and/or what are the inter-dependencies among them.

Table 3. Reflections of group work and discussions about the dimensions in QoS framework

DIMENSION	WHAT TO COVER IN ASSESSMENT
Resources and their management	
Human resources	• researcher capacity in terms of qualifications and scientific track record • performance management and incentive system • ratio between senior and junior scientists, mentoring • research teams: disciplinary balance and overall composition • research leaders (H index, CV, leadership capacity) • research partners (reputation, capacity of researchers) • publication productivity
Financial resources	• influence of financial resources on QoS (maintaining critical mass, ability to attract co-funding, influence of funding fluctuations on QoS)
Infrastructure	• laboratories, experimental facilities and their use
Research process	
Research planning and design	• processes for research planning, appraisal, peer review • assuring quality of ToC at Flagship level through use and updating • methodological quality and credibility of design • mechanisms for quality assurance at planning stage • ethical/institutional review
Research implementation and protocols	• research process state-of-the-art, efficiency of research process, synthesis and knowledge accumulation • benchmarking e.g. with private sector (plant breeding) • data management systems and quality control
Outputs	
Scientific and peer review	• traditional metrics used in diagnostic manner (not for ranking)
Non peer reviewed/written with science reporting purpose	• scientific quality plus utility and potential impact in the scientific community
Non publication output	• output-specific criteria, for example novelty or level of trait targeted in improved germplasm; use of modules, tools and portals

4. MAIN ISSUES DISCUSSED

This section summarizes the main issues that emerged during the discussion following the presentations, as well as during the panel and group discussions.

Context and scope

Definition of Quality of Science

Major discussion centered on the fundamental issue of defining QoS, in addition to what is to be assessed within this criterion. As noted earlier, there is currently no commonly applied definition of QoS within the CGIAR.

The difference between QoS and *quality of research* was discussed. Some workshop participants were of the opinion that quality of research as a term would be more useful in evaluation since it is broader and includes the process (selection of research topics and modalities of conducting research), and many of the elements in the evaluation framework and proxy indicators cover a broader area than QoS in a strict sense. While QoS is at the core of each CRP, the evaluation needs to cover also the entire research process that involves these other aspects related to the research process. Research-related activities such as research capacity development and knowledge management that do not produce science but support the development orientation of research also need to meet quality standard and be evaluated for that.

In conclusion, the participants agreed that the term Quality of Science should be kept, even if its evaluation is expanded to include a broader set of research process-related aspects, for its particular relevance to CGIAR work:

- QoS allows a better framing of the assessment framework since it is a component requirement of overall research, research being the entire package in CRPs that are programs of research;
- QoS is an established term in the CGIAR and carries particular significance, and rather than replacing it, we need to explain it better.

A view was presented that even in the broad context of evaluating research quality involving research process, it should be possible to define what is pursued through the various indicators in determining the *quality* of science (as the knowledge produced). It was suggested that this relates to the robustness of the knowledge/research findings in reflecting the causality or phenomena studied – that the findings are not artifacts. Clarity about the essential property defined as quality in science would be helpful for the broader analysis of both quality and relevance as essential components of program performance. With this in view, there is need to adjust the QoS assessment framework to include elements of research process currently not covered and, in the overall evaluation framework and guidance to evaluators, specify the relationship between quality and relevance as well as the dependency of effectiveness on quality and relevance.

Quality of Science in relation to other evaluation criteria

Relating to the above discussion on scope is the issue on how the QoS criterion links to the other evaluation criteria. Moving from research outputs based on high quality science to outcomes is a key question for overall research performance assessment.

In terms of evaluation criteria, it was noted that "Relevance" has a particularly close relationship to QoS. Participants generally agreed that the assessment of relevance should be integrated with the assessment of QoS since achievement of desired outcomes depends on both criteria being met.

The aspects of relevance that are particularly closely related to the QoS and the research process are:

- Relevance to the CGIAR System Level Outcomes and comparative advantage
- Relevance to the research community
- Relevance and quality of the engagement with user and beneficiary communities; i.e. the legitimacy of the research

The challenge is to put all the pieces of the puzzle together and to combine the assessment of the various criteria. For several CRP evaluations, evaluation teams have assigned one team member the responsibility for a chapter on QoS. There has been a tendency of presenting conclusions and recommendations on QoS in isolation, without addressing linkages to other components of the evaluation and how the strengths and limitations in any component affect the overall program performance. While the teams have not assess research efficiency in terms of cost-effectiveness, there are aspects of the efficiency of the research process that have been included in the QoS assessment. An example of that is the state-of-the-art in a breeding program. It is therefore important that (i) QoS is addressed within the context of the overall program assessment, tailoring the assessment appropriately to the substance of research; and (ii) that the implications of QoS findings are related to other findings. The boundaries of the evaluation criteria in gathering and assessing evaluative information and structuring of the evaluation report by criteria should not restrict cross-criteria analysis.

Quality of science in the CGIAR context

There was discussion around the expectations with regard to QoS, especially since CGIAR works on research for development and along the spectrum between research and development, and is increasingly being held accountable for performance along the impact pathway towards outcomes. While it was agreed that not all research projects have to be scientifically at the leading edge, research anywhere along the spectrum should be of high quality producing scientifically credible results. In addition, the CGIAR has a role to bridge between knowledge creation and knowledge utilization. The expectation of high quality also applies to locally relevant research. For example, applied research done in participatory manner with non-scientists also needs to meet with scientific standards that assure the validity of the findings. The general agreement from all participants was that QoS is of high importance, and that it is a prerequisite for achieving development outcomes and needs to be an explicit component of research evaluations.

The expectation for development application and impacts also influences research processes and QoS. The CGIAR is unique in that it addresses fundamental issues of agricultural development through research, which very few other agencies can do. Some research has very long and protracted impact pathways and, due to the inherent risk in research, may not produce expected results even if, during the process of investigation, valuable and publishable research findings may be generated. One suggestion was that even when we do not know whether a research initiative will produce results and whether these results eventually "take off" we should be able to say at what point we should know. This also includes having stop/go points during the research process at which managers have to make strategic decisions about continuation of research. A cautionary note on this was that the CGIAR needs to be

careful in applying “business like” approaches, since in many cases research is not about a linear accumulation of knowledge within strict timelines.

Additional points raised included:

- CGIAR research cannot be only locally relevant, and needs to be globally relevant, as the ability to generate IPGs is part of the comparative advantage of CGIAR; thus the IPG nature of research is an aspect of QoS evaluation;
- There is a need to distinguish between applied, strategic and highly exploratory research (with potentially very long term and high risk impact pathways) in designing QoS assessment and how it is linked to evaluating the overall performance of research;
- In linking QoS evaluation to that of overall program relevance and performance, the relevance of science outputs becomes inseparable from the QoS assessment; otherwise high scientific publications and citation record may hide research that has little relevance to the issues addressed by the CRP and its objectives;
- It is important that in assessing excellence in science the standards for QoS are not, due to the mission-orientation of the CGIAR, set lower than elsewhere.

Responsibility for science quality (Centers versus CRP)

An underlying, and unresolved, issue is the ambiguity relating to Centers versus CRPs when it comes to responsibilities regarding science quality and evaluating research management. With multiple Centers implementing CRPs, the CGIAR has a matrix management structure that affects also how QoS is managed and overseen. The legal entities in the CGIAR are the Centers; and they are the employers of the researchers and owners of laboratories and each have top research management and Board of Trustees overseeing research. In many cases, a CRP Director does not have adequate authority over the inputs and processes which would affect the QoS in the program.

From an evaluation point of view, the fact that the teams are evaluating CRPs, while some accountability remains with Centers, makes it difficult to identify the causes of shortcomings in respect of QoS and target recommendations at those accountable. Occasionally, CRP records disguise the Center origin or contribution of research results. Conversely, it is often not clear as to which scientists contribute to a CRP and how much, and which published Since Centers make a commitment to the implementation of CRPs, they are expected to account for the quality of their research contribution. Questions such as “What can the CRP management and oversight do to have a better handle on the QoS?” and “How does the CRP use its authority to demand more from the Centers?” are important in that regard, and should be key questions to address in the evaluations.

Quality of Science within the Theory of Change

The ToC can provide a useful framework for analyzing QoS in the context of the broader program performance. Most donors/CGIAR investors don’t view the success of their investments solely in terms of QoS. A view emerged that QoS is not an end in itself, although a necessary condition, and that it is equally important to look at the value of the science, which can be assessed from the potential and actual contribution to progress along the TOC on one hand and to updating the TOC as a dynamic planning tool, on the other. Furthermore, it was noted that the programs ToCs, and research design have to link to the overall CGIAR Strategy and Results Framework, and therefore evaluations need to consider QoS in that context and in combination with the rest of the evaluation. AAS evaluation was

highlighted as an example of one that successfully linked QoS with the assessment of the overall research design and ToC.

Methodology

The framework

There was a general sense that the IEA framework, while being very comprehensive, runs the risk of being too cumbersome to apply, as it does not prioritize among the QoS aspects or consider the relative strength of different indicators. . However, it was also acknowledged that by trying to achieve more specificity and consistency, which often implies using quantitative data and indicators, the risk exists that we sacrifice expertise in providing nuances for analysis. Ultimately, there needs to be a balance between being comprehensive yet not overly prescriptive, and distilling the essential messages about QoS. In the working groups, suggestions were made on how to modify the components of the framework, and additional criteria, for example legitimacy (the extent to which CGIAR research outputs are perceived trustworthy and represent the interests and values of the intended users) were proposed. However, the participants did not find a way of reducing or compressing the quite large number of aspects that eventually were suggested to be included in the framework (see Table 3 above).

Participants noted that the framework assumes continuity between input, management processes and outputs. However, the outputs of the CRP are assessed quite separately through bibliometric analysis and peer review of sample publications or other outputs. Consequently, evaluations have seldom provided an explanation to why outputs are of good or poor quality, or assessed the implications of the quality of input and management processes on outputs. The difficulty in analyzing these dependencies relates to the fact that multiple Centers provide the inputs, management processes and outputs and are responsible for their quality. Tracking the linkages Center by Center is challenging.

Using Quantitative and qualitative data

The IEA framework uses a mixed method approach to assessing QoS. The evaluation methods used have been partly quantitative (bibliometric analysis, scoring of research outputs, researcher survey) and partly qualitative (expert review, interviews, etc), and interpretation of the data and information has been through a qualitative team process. Interestingly the experience from previous evaluations shows that the greatest variability in conclusions actually related to the quantitative assessments, due to different team interpretations.

In the discussion, the challenges related to indicators and scores included:

- Indicators often capture only quantity rather than quality (e.g. number of publications);
- When using scores, experts tend to focus only on the score and do not contextualize their analysis;
- Linking different indicator values and scores together for a higher level assessment can be difficult, indicator information may be too limited for illustrating the bigger question of performance;
- Scoring within disciplinary boundaries may dismiss elements of the interdisciplinarity that have their own quality issues;
- Indicators related to outputs do not address the relevance and usability of them;

- The interpretation of indicators that is necessary in evaluation can vary considerably (see discussion below).

However, it was generally agreed that quantitative indicators have their value but their results should be triangulated with other evidence, and that IEA should continue to use both qualitative and quantitative means when assessing QoS. In the case of AAS for example, the evaluation captured serious concerns around staff resources by triangulating data on distribution of researchers with other evidence. On the other hand, case studies often combine qualitative and quantitative information, and for example in the L&F evaluation, case studies proved to be very useful for better understanding the whole research process from design to implementation to output.

As one participant put it: "We tried to use multiple methods wherever we go, wanted to get as high a degree of triangulation as possible. Using mixed methods was a way to get consistency across the team. We were trying to generate multiple narratives that give you greater confidence that what you see is real and where you should dig deeper".

Interpretation of evidence

One of the major lessons learned from the completed CRP evaluations was that there are considerable differences in the interpretation of evidence and in several cases the assessment did not go beyond rankings or descriptive statements. In this context, the following points and issues were noted:

- Benchmarking: who do evaluators benchmark their findings with; who does CGIAR compare itself with?
- Recognition that there are no benchmarks for some areas; perhaps guidance is needed to specify where the evaluators could look for clear internal or external benchmarks and where experts need to benchmark against their expectation;
- Subject matter specialists may be hesitant to assess their peers or they may set their expectations across whole program according to their specific area of expertise;
- Indicators can indicate more than one thing and looking at them in isolation may be misleading;
- It is very difficult to force consistency in the interpretation of evidence nor should it be the main priority, evaluation teams need flexibility;
- Individual evaluators may conduct their assessments at different levels of detail that does not necessarily mean disagreement, but needs to be resolved in evaluation conclusions.

It was generally acknowledged that there is a human factor in every assessment, and a level of variability needs to be accepted. It can, however, be reduced by having a better common understanding on key issues and what we understand by QoS.

Research design – an important aspect

It was felt that research design has not been fully addressed in the evaluations despite its importance. Only the AAS evaluation looked closely at the research design from a QoS perspective, in addition to analyzing the likelihood of impact and whether research was lined up to deliver on the promised outcomes. Research design at project level is concerned with issues such as formulation of hypothesis and research questions, choice of methods and sampling and extrapolation of results. To assess these details, one needs project specific information and documentation, which, in many cases, was not available to the evaluation teams.

There was a view that research design can be looked at at different levels. There is a hierarchy of research design starting from project level and going to overall CRP level. The analysis would need to cut across the different levels and eventually assess the coherence of components and the overall program design. An issue is that often an overall programmatic research design does not exist, especially if many programmes consist largely of bilaterally funded projects. These important aspects span across the QoS and relevance criteria, which needs to be taken into account in the assessment.

5. CONCLUSIONS FOR THE FUTURE

The workshop was a first step to reflect on IEA's experience in recent evaluations. The discussion covered, on the one hand, the multi-dimensional framework required for assessing QoS. On the other hand, it also highlighted the challenges of prioritizing the indicators and information required within that framework and the complexity of translating the framework into an operational assessment tool that would enhance consistency and comparability.

The IEA's responsibility is to conduct high quality evaluations that address QoS in the most appropriate way as one criterion among other reflecting overall program performance. In order to capitalise on the rich experience in the completed evaluations, and enhance the quality, consistency and relevance of the future evaluation of research by IEA and in other parts of the CGIAR, the IEA plans to develop guidelines for assessing QoS when evaluating research programmes. This will need to be done in concert with other discussions on QoS and in close consultation, particularly with the ISPC but also drawing from expertise and perspectives in and outside the CGIAR.

QoS issues touch upon all CGIAR functions of appraisal, monitoring and evaluation of research and assessment of its impact. As the three System entities of the CGIAR (Consortium Office, ISPC and IEA) thus all address QoS issues and inform the same stakeholders (Centers, CRPs, donors and partners), enhancing common understanding of QoS among the entities, including its definition, is important. However, scientists and research managers involved in research are producers of science and their involvement in discussing science quality definition and principles is essential. Furthermore, as the work of CGIAR scientists is the subject of evaluations, the understanding of what is expected of research programs and what they will be evaluated against need to be shared.

The IEA concludes that the workshop, through its rich discussion and ideas brought up, fulfilled the expectations and was helpful for the IEA to improve research evaluations with regard to assessing QoS. The IEA will move forward in preparing guidance for future evaluations and in doing so, consult with others with responsibilities over QoS. Through its analysis of evaluation experiences, the IEA also intends to contribute to the broader discussion on QoS led by the ISPC.

ANNEX A: Participant list

Name	Position	Organization	Contact
EXTERNAL			
Brian Belcher	Senior Associate Scientist	CIFOR	B.Belcher@cgiar.org brian.belcher@royalroads.ca
Regina Birner	Chair of Social and Institutional Change in Agricultural Development	University of Hohenheim	regina.birner@uni-hohenheim.de
Marlene Diekmann	Research Advisor	GIZ Advisory Service on Agricultural Research for Development	marlene.diekmann@giz.de
Philippe Ellul	Senior Science Officer	CGIAR Consortium	p.ellul@cgiar.org
Julie Fitzpatrick	Director	Moredun Research Institute	Julie.Fitzpatrick@moredun.ac.uk
Margaret Gill	Chair	ISPC	m.gill@abdn.ac.uk
Tim Kelley	Senior Agricultural Research Officer	SPIA	Timothy.Kelley@fao.org
Christian Roth*	Research Program Manager	CSIRO	Christian.Roth@csiro.au
Rachid Serraj	Senior Agricultural Research Officer	ISPC	Rachid.Serraj@fao.org
Tricia Wind	Senior Program Officer Corporate Strategy and Evaluation Division	IDRC	twind@idrc.ca
IEA			
Rachel Bedouin	Head	IEA	Rachel.Bedouin@fao.org
Manuela Bucciarelli	Evaluation Analyst	IEA	Manuela.Bucciarelli@fao.org
Federica Coccia	Evaluation Analyst	IEA	Federica.Coccia@fao.org
Sirkka Immonen	Senior Evaluation Officer	IEA	Sirkka.Immonen@fao.org
Sophie Zimm	Evaluation Analyst	IEA	Sophie.zimm@fao.org
Jenin Assaf	Communications Officer	IEA	jenin.assaf@fao.org

ANNEX B: Detailed agenda

	Day 1 (THURSDAY, 10th December) Setting the context and reflecting on other experiences	Responsibility
9:00 - 10:00	**SESSION 1: Setting the context – Assessment of Quality of Science (QoS) in IEA evaluations**	**Chair:** **Rachel Sauvinet-Bedouin**
9:00 – 9:15	Introduction to the workshop and to Agenda Introduction of participants	**Rachel Sauvinet-Bedouin**
9:15 – 10:00	Presentation of the IEA QoS evaluation framework - *Background to the development of the IEA's QoS evaluation framework* - *What the QoS evaluation framework covers* - *Scope of QoS and link to the other evaluation criteria* Questions for clarification (and raising issues for further consideration during workshop)	**Sirkka Immonen** **All**
10:00 – 12:00	**SESSION 2: Learning from IEA experiences**	**Chair: Regina Birner**
10:00-10:30	Presentation of how QoS was assessed in IEA evaluations - *Methods used across evaluations* - *Programme specific adjustments in approach* Questions for clarification (and raising issues for further consideration during workshop)	**Sophie Zimm** **All**
10:30 – 10:45	COFFEE BREAK	
10:45-11:15	Lessons learned from CRP evaluation synthesis - *Patterns of findings and recommendations relating to QoS* - *Lessons from interpretation of evidence* - *How was QoS linked to overall CRP evaluation?* Questions for clarification (and raising issues for further consideration during workshop)	**Christian Roth (skype)** **All**
11:15 - 12:15	Panel discussion – experiences from IEA evaluations - on designing the QoS framework for evaluations (from ToR to Inception Report) - on implementing the framework (data collection, analysis, interpretation, etc)	**Evaluation team members and managers**
12:15-13:15	LUNCH	
13:15-15:00	**SESSION 3: Learning from other experiences on QoS assessment**	**Chair:** **Tim Kelley**
13:15-13:45	Presentation: ISPC assessment of QoS in proposal appraisals	**Maggie Gill**

	Questions for clarification (and raising issues for further consideration during workshop)	**All**
13:45-14:15	Presentation: IDRC experience Questions for clarification (and raising issues for further consideration during workshop)	**Tricia Wind** **All**
14:15-15:00	Discussion on IEA evaluation in light of other experiences - *What are the main issues from IEA experience to-date (IEA, team and user perspectives)?* - *What lessons from outside the IEA are most relevant to be brought to the discussion?*	**All**
15:00 – 17:30	**SESSION 4: Working groups**	
15:00 – 17:30 Coffee break included	**Working Groups** Introduction to Working Group assignment Breakout groups to discuss the three dimensions of IEA framework (inputs, management, outputs)	**All**
20:00	DINNER	
Day 2 (FRIDAY, 11th December) Looking forward and planning for the future		**Responsibility**
9:00-10:30	**SESSION 4 CONTINUED: Working group results**	**Chair: Brian Belcher**
9:00 – 9:40	Reporting back from working groups (around 10 min each) - *Ideas coming back from groups*	
9:40-10:30	Discussion: Bringing different QoS dimensions together	
10:30-11:00	COFFEE BREAK	
11:00 – 12:30	**SESSION 5: Conclusions and way forward**	**Chair: Sirkka Immonen**
11:00 – 12:30	Conclusions for strengthening evaluation of QoS for improving its usefulness - *What should evaluation of QoS include and what areas should be strengthened to help decision making for improving QoS?* - *Developing IEA Guidelines for improving QoS evaluation: evidence, interpretation and linking findings to overall evaluation (what should they cover, which issues should they address, how prescriptive should it be)* - *Linking IEA QoS evaluation to the broader discussion of QoS in the CGIAR* - *Closure*	**Rachel Sauvinet-Bedouin**